LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-687422

Characterization of genetic variability of Venezuelan equine encephalitis viruses

S. Gardner, K. McLoughlin, N. Be, J. Allen, S. Weaver, N. Forrester, M. Guerbois, C. Jaing

March 31, 2016

Public Library of Science

Disclaimer

Characterization of genetic variability of Venezuelan equine encephalitis viruses

Shea N. Gardner[1], Kevin McLoughlin[1], Nicholas A. Be[2], Jonathan Allen[1], Scott C. Weaver[3], Naomi Forrester[3], Mathilde Guerbois[3], Crystal Jaing *

[1]Computations, Lawrence Livermore National Laboratory, Livermore, California

[2]Physical and Life Sciences, Lawrence Livermore National Laboratory, Livermore, California

[3]Institute for Human Infections and Immunity and Departments of Microbiology & Immunology and Pathology, University of Texas, Medical Branch, Galveston, Texas

*Corresponding author

Email: jaing2 ć llnl.gov (CJ)

LLNL-JRNL-687422

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Abstract

Venezuelan equine encephalitis virus (VEEV) is a mosquito-borne alphavirus that has caused large outbreaks of severe illness in both horses and humans. New approaches are needed to rapidly infer the origin of a newly discovered VEEV strain, estimate its equine amplification and resultant epidemic potential, and predict human virulence phenotype. We performed whole genome single nucleotide polymorphism (SNP) analysis of all available VEE antigenic complex genomes, verified that a SNP-based phylogeny accurately captured the features of a phylogenetic tree based on multiple sequence alignment, and developed a high resolution genome-wide SNP microarray. We used the microarray to analyze a broad panel of VEEV isolates, found excellent concordance between array- and sequence-based SNP calls, genotyped unsequenced isolates, and placed them on a phylogeny with sequenced genomes. The microarray successfully genotyped VEEV directly from tissue samples of an infected mouse, bypassing the need for viral isolation, culture and genomic sequencing. Finally, we identified genomic variants associated with serotypes and host species, revealing a complex relationship between genotype and phenotype.

Keywords

Venezuelan equine encephalitis virus; microarray; single nucleotide polymorphism; genotype; phenotype; pathogen

Introduction

Venezuelan equine encephalitis (VEE) virus (VEEV) is a mosquito-borne alphavirus capable of causing large outbreaks of encephalitis in humans and horses. Major equine-amplified epidemics dating to the early 20th century have affected hundreds-of-thousands of people and economically important equids. VEE complex viruses are endemic to South and Central America, Mexico, and Florida [1]. Although the case-fatality rate of VEEV is low in human infections (usually less than 1%), infection is typically highly debilitating and sometimes results in permanent neurological sequelae [2]. Moreover, because the disease primarily occurs in isolated rural areas and typical infections initially present with nonspecific flu-like symptoms, many cases involving spillover from enzootic cycles go undiagnosed or are mistaken for other febrile diseases such as dengue [3].

Enzootic VEE is also of concern due to its high burden of endemic human disease. For U.S. war fighters engaged in a conflict in Latin America, either direct exposure to the enzootic cycle in rural or suburban regions, as documented in Panama [2, 4], Colombia [5], and Mexico, or infections in urban settings [3, 6-10] could inflict direct casualties and severely compromise their ability to fight.

There are three major challenges related to VEE that we believe can be solved using new approaches: 1) rapidly estimating the origin of a newly discovered VEEV strain; 2) estimating its equine and/or human amplification, and thus epidemic potential; and 3) predicting the human virulence phenotype of a newly discovered VEEV strain. Phylogenetic relationships of a diverse collection of VEEV strains have proved useful for identification of the genetic features leading to epidemic spread to humans and livestock of this zoonotic pathogen [11, 12]. Here, we exploit high-throughput technologies to characterize a large panel of strains, including both virulent and

avirulent strains; geographically diverse isolates from South America, Central America, Mexico, Florida and Texas; and isolates of multiple serotypes from diverse hosts, including human outbreak strains.

We performed whole genome SNP analysis of all available VEE antigenic complex genomes, verified that these SNPs accurately recapitulated the phylogeny from whole genome multiple sequence alignment (MSA), and developed a high-resolution genome-wide SNP microarray. We analyzed a diverse panel of 133 VEEV isolates on the microarray to validate array-based SNP calls with previously sequenced strains, and to characterize the SNPs in unsequenced isolates and place them on a phylogeny with sequenced genomes. We explored the relationship between genome variation and serotype, identified a number of variants non-randomly associated with these phenotypes, and examined the distribution of these variants across the VEEV genome.

Methods

Whole genome SNP analysis

We applied the kSNP software to find SNPs in the 144 VEE antigenic complex genomes available as of June, 2014 [13, 14]. kSNP is an alignment-free method based on examination of *k*-mers (oligos of length *k*) in the genome sequences. We define a SNP locus by a sequence context of length *k* centered on the polymorphic base, with (*k*-1)/2 conserved bases on either side. For this study, we performed SNP analysis with *k*=13. Note that, under this definition of SNP loci, multiple loci (corresponding to different variations of the *k*-mer context) may overlap

the same positions in a multiple sequence alignment; in this case, each of the multiple loci is only considered to be present in the genomes in which the (k-1) base context is conserved. This alignment-free SNP discovery is useful for viruses in which there may be highly divergent and poorly alignable regions among a large group of sequences, and where conserved regions only exist among small subgroups of sequences. The kSNP approach is free of the bias that otherwise results from the choice of a reference sequence, or from considering only a subset of regions of the genome that can be easily aligned, and can be implemented at scales to hundreds of genomes. We calculated SNP-based phylogenetic trees using parsimony, maximum likelihood (ML), or neighbor joining (NJ). For NJ, we used the number of SNP allele differences between pairs of target sequences as the distance metric. We mapped SNP alleles to branches of the trees using kSNP.

Tree comparisons

SNPs from the E1, E2, E3, and capsid genes were extracted for separate analyses by identifying those SNPs that occurred within the specified gene regions (Table 1). We constructed a full genome MSA using the MUSCLE software [15], and built parsimony trees from the MSA, from all SNPs, and from SNPs in each gene. We compared the MSA and gene-based trees to the all-SNPs tree by treating all trees as unrooted and examining the splits of isolates into pairs of groups on either side of each internal branch in the tree. For each tree we used the Perl script CompareTree.pl [16] to calculate the fraction of splits shared between the MSA or gene-based tree and the all-SNPs tree; this serves as a metric of similarity between the tree topologies. We also used Dendroscope [17] to generate tanglegrams, which display pairs of trees side by side with lines interconnecting corresponding taxa. To minimize the numbers of crossing lines

between trees without changing the tree topologies, we performed a series of equivalent branch rotations using the algorithm in [18] before generating tanglegrams. The pattern of crossing lines remaining provides a direct visualization of the differences in tree structures.

Table 1. Gene regions from which SNPs were extracted.

Gene	Coordinates on TC-83 genome
E1	10000-11327
E2	8563-9843
E3	8386-8574
Capsid	7562-8396
All SNPs	1-11446

Microarray probe design

We designed microarray probes for every SNP locus. Our probe design strategy maximized sensitivity and specificity based on extensive prior lab testing on a Roche NimbleGen microarray platform, where we demonstrated 99.52% SNP allele call rates and 99.86% accuracy [19]. After testing seven alternative probe design strategies, we determined that maximum sensitivity and SNP discrimination accuracy resulted if the SNP base was at the 13^{th} position from the 5' end of the probe (the end farthest from the array surface), probes were between 32 and 40 bases long, and lengths were chosen to equalize hybridization free energy (ΔG) to the extent possible within the allowable length range. We found that probes shorter than 32 bases had high false negative rates, and longer probes did not discriminate well between alleles. We found that ΔG was a better predictor of hybridization than the melting temperature T_m. Probe candidates with hybridization free energy below ΔG_{min} = -43 kcal/mol were shortened until either

their free energy exceeded ΔG_{min} or they reached the minimum 32 bases. Probes were designed around the SNP on both the plus and minus strands, for all observed SNP alleles, and all surrounding sequence variants.

Probes for the plus and minus strands were not the reverse complements of one another because the SNP does not lie at the center of the probe. We included probes for all observed alleles on each strand, yielding at least four probes per SNP locus for biallelic SNPs. In addition, we captured any sequence variation outside of the conserved k-mer SNP context in multiple alternative probes for each allele, so that some biallelic loci had more than 4 probes. Finally, we trimmed probes from the 3' end to remove any N's or other ambiguous bases, and omitted them altogether if doing so resulted in a probe shorter than 32 bases. When a probe was a subsequence of any other, only the shorter of the two was kept. SNP microarrays were fabricated using the 12-plex 135K Roche NimbleGen array format with 89% of the probes tiled in duplicate.

Array hybridization to VEEV cDNA samples

The VEEV cDNA samples were fluorescently labeled and hybridized to VEEV SNP arrays as described previously [20]. Briefly, fluorescent labeling of samples was performed using the NimbleGen One-Color DNA Labeling Kit (Roche). One μg VEEV cDNA was added to Cy-3 labeled random primers, followed by isothermal amplification at 37°C using Klenow polymerase. Labeled DNA was purified via isopropanol precipitation and resuspended in water for microarray hybridization. DNA samples were prepared for hybridization using the NimbleGen Hybridization Kit LS (Roche). Three μg of labeled DNA was hybridized to each array, incubating for 40-45 hours at 42°C. Arrays were washed using the NimbleGen Wash Buffer Kit (Roche). The fluorescent signal on the array was scanned using a 2 μm Roche MS200

fluorescent scanner. Array feature intensities were generated using the NimbleScan software available from Roche NimbleGen.

Selection of VEEV isolates for microarray experiments

Based on temporal and geographic range, outbreak associations and prior sequences generated at UTMB, we identified, propagated, and isolated RNA for microarray experiments from 134 of the most representative strains. To enable comparison of array- and sequencing-based genotyping methods, we included 81 isolates that had previously been sequenced in the set of strains tested on the array. Three of the previously sequenced isolates and one unsequenced isolate were run on duplicate arrays, for a total of 138 arrays. The serotype, passage history, year and location of collection, and host of each strain are listed in S1 Table.

To test the array's ability to genotype viruses directly from tissue samples, six day-old CD-1 mice were infected with VEEV vaccine strain TC-83 [21] via the intracranial route. Each mouse was infected with 10^4 PFU in a 20 μL volume. Three biological replicate mice were infected and sampled. Brains were harvested two days later and homogenized in a 1:10 w/v solution. The suspension was clarified by centrifugation and stabilized in Trizol (Life Technologies). RNA was extracted and purified using the Direct-zol RNA MiniPrep kit (Zymo Research, Irvine, CA) according to the manufacturer's instructions. Whole cDNA was synthesized, fluorescently labeled, and hybridized to the SNP microarray as described above.

Data from microarray experiments is available at the Gene Expression Omnibus (GEO) repository under accession GSE79530.

Allele calling from SNP microarray data and concordance calculations

We used our previously developed analysis software to call alleles at each locus for each sample analyzed on SNP microarrays. The software fits a linear model of strand and allele effects to the log intensity data from all probes for the locus, and calls the allele as the one with the largest coefficient in the fitted model. Separating the strand and allele effects is necessary in order to compensate for the differing hybridization efficiencies often seen between forward and reverse strand probes.

Because our definition of a SNP locus requires conservation of the 6 bases on either side of the polymorphic base, array probes for one locus may hybridize to genomes in which a similar locus context is present. That is, loci that are considered to be different in the sequence analysis, but have 13-mer contexts that are identical except at one or two positions, may be difficult to distinguish by microarray probes. Therefore, our current array analysis software does not attempt to determine whether a locus is present or absent, and instead makes an allele call for every locus.

For isolates that had genome sequences available, we computed the concordance rate between the allele calls from the array and the genome sequence, defined as the percentage of loci present in the genome for which the array calls agreed. We also computed the numbers of allele differences between each array sample and each genome, and determined whether the closest genome was in fact the genome sequence for that strain.

Analysis of phylogenetic relationships and evolution of VEEV strains from SNP microarray data

We used the genotype data from genomic sequences to create maximum parsimony phylogenetic trees, using Parsimonator (https://github.com/stamatak/Parsimonator-1.0.2). We generated 100 trees using different random number seeds, and selected the most parsimonious (that is, the tree requiring the smallest total number of nucleotide substitutions) for downstream analyses.

Phenotype/genotype associations

We identified variable positions in the MSA and used these loci as an initial set for building decision tree classifiers, using the recursive partitioning algorithm implemented in the R function "rpart" from the package "mvpart" [22, 23]. The "rpart" algorithm is described in detail in [24]; briefly, it selects a series of variables (SNP loci) and values (alleles) that split the viral strains into groups with homogeneous phenotypes. Each split of a group into smaller subgroups is chosen to minimize the Gini index, a measure of total subgroup inhomogeneity.

To ensure there were sufficient samples in the training and test sets for each phenotype to be predicted, we defined a "host type" for each sample by categorizing hosts as "large" (humans and equids) or "small" (rodents and mosquitos). For each phenotype (serotype and host type), we built multiple tree classifiers using a 10-fold cross-validation scheme, in which classifiers were trained with 90% of the isolates and tested with the remaining 10%. The amount of pruning in each decision tree classifier was determined by a complexity parameter; the "rpart" algorithm automatically determined an optimal complexity, defined as the smallest parameter value that yielded a cross-validation error rate within one standard deviation of the minimum error rate. We then built a final decision tree for each phenotype with the full set of genomes, using the optimal parameter to control the complexity of the tree. For each phenotype, we tested initial locus sets consisting of all polymorphic loci present in the TC-83 reference genome, as well as restricted

sets containing only non-synonymous loci within the genes encoding structural proteins, or within the envelope glycoprotein genes.

For each classifier, we computed an overall accuracy, defined as the percentage of phenotype predictions that were correct. We also computed performance metrics for each specific phenotype, treating the decision tree as a binary classifier; e.g. for classifying isolates as serotype IAB vs any other serotype. For each specific phenotype, we counted the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), and used them to compute the accuracy = (TP + TN) / (TP + FP + TN + FN), positive predictive value (PPV) = TP/(TP + FP), negative predictive value (NPV) = TN/(TN+FN), true positive rate TPR=TP/(TP+FN), and true negative rate TNR=TN/ (TN+FP).

We also ranked loci according to their strength of association with serotype or host type according to Fisher's exact test, as implemented in the R "fisher.test" function. We corrected p-values for multiple comparisons using the Benjamini-Hochberg method.

Results

Whole VEEV genome SNP analysis

To identify single nucleotide variations among VEEV strains, we applied the kSNP software to 144 VEE antigenic complex genomes. We identified 7,926 SNP loci among these strains. The numbers of SNPs identified in structural protein encoding regions are summarized in Table 2. The annotations, 13-mer contexts and reference genome alignments for SNPs identified by whole genome analysis are listed in S2 Table.

When we reran the kSNP analysis, using as an outgroup four strains of eastern equine encephalitis virus (EEEV, the closest relative of VEE complex alphaviruses), the total number of SNP loci increased to 9,486.

Table 2. Numbers of SNPs identified in VEEV genomes, by gene region.

Gene	Number of SNPs
E1	1268
E2	1384
E3	262
Capsid	937

Phylogenetic tree construction

We then examined phylogenetic relationships among strains by building trees using different methods. First, we wished to determine which of the SNP-based tree construction methods performed best, by comparing the resulting trees to trees based on whole genome multiple sequence alignment (MSA). We found that the SNP tree built using maximum parsimony (Fig 1) was more similar to the MSA-based tree than those built with NJ or ML. Out of all splits in the alignment-based tree, 77% were also present in the parsimony tree, compared to only 68% in the ML tree. Moreover, the parsimony tree had fewer homoplastic SNPs than the ML tree (1679 versus 2153, respectively, from the dataset using the EEEV outgroup genomes). Homoplastic SNP loci are those in which the pattern of shared alleles does not conform to any of the branches of this tree, as a result of processes such as convergent evolution, homologous recombination, multiple mutations at the same site, or sequencing errors. Maximum parsimony has been shown to outperform ML in phylogenies that display heterotachy, a phenomenon in which the rates at which different nucleotide positions evolve change over time [25]. In this case,

non-parametric estimation of trees by parsimony is more accurate than parametric methods such as ML.

Almost all VEEV strains could be uniquely identified by their genotypes according to variations across the identified SNP loci. Numbers at the interior nodes of the tree in Fig 1 indicate the number of loci at which a SNP allele is uniquely found in the descendants of the node and is shared by all of them. Only two sets of genomes were unresolved (i.e., had identical genotypes across all 7,926 SNPs); these strains are labeled in Fig. 1 with italic type. One consisted of two genomes collected on successive days from Minatitlan, Mexico on August 26-27, 2010: MX10_91M8 from a mosquito pool and MX10H91_00011 from a sentinel hamster. The other comprised four genomes, also collected from Minatitlan in 2010; MX10_94M4, MX10_94M5 and MX10_94M6, collected from mosquito pools on August 26-27, and MX10H95_00014, collected from a hamster on August 28. These results confirm that sentinel hamsters do become infected with the variants circulating in mosquito vectors in the area at the same time. These isolates were members of a larger group of closely related genomes collected in Minatitlan, Mexico between July 2008 and late August 2010 from hamsters, mosquitos and two horses.

Phylogenetic and phenotypic relationships of VEEV strains

To explore the relationship between the phylogenetic groupings of VEEV strains and their phenotypes, we examined the maximum parsimony tree shown in Fig 1, in which the genome annotations and plot symbols are color-coded by serotype and host, respectively. We observed a number of interesting patterns. First, we extended previous results [26] showing that VEEV strains with high overall similarity across the entire genome may exhibit different serotypes. For example, the epizootic serotype IAB strains and associated vaccine strain TC-83

(purple in Fig 1) collected from multiple countries from 1938-1973 form a distinct clade of highly similar isolates; however, this clade also included a serotype ID isolate (R16905) collected in 1977. In general, we saw that broad phylogenetic groupings were not exclusively associated with particular serotypes.

Similarly, we found that phylogenetic groupings were not strongly associated with particular hosts; the broad associations that did appear were likely artifacts of the different sampling strategies used for enzootic (serotype ID and IE) strains, which account for all samples from mosquitos and sentinel hamsters, and for epizootic (serotype IAB and IC) strains, which comprise most samples from equids and humans.

Finally, when we examined the collection dates of samples found in each major clade, we found that many clades were remarkably persistent. For example, the serotype IAB epizootic strains (and associated type ID outlier) showed little genetic variation, even though they were collected over nearly 40 years (1938-1977) across a wide geographic area, from the USA through Guatemala and Trinidad down to Venezuela and Peru, likely the result of incompletely inactivated vaccines made from older strains initiating later outbreaks [27]. Likewise, the serotype IC and ID isolates comprising the lower part of the tree in Fig 1, collected between 1961 and 2005, had very few differences across our panel of SNP loci.

Association between genotypes and phenotypes

Because the host and serotype associated with a VEEV isolate are not completely predictable from its position in the phylogeny, we searched for SNP loci that were associated with these important phenotypes for which the association was not simply a product of ancestry. We applied the "rpart" recursive partitioning algorithm to identify variations that are associated

with particular host types or serotypes. The resulting decision tree classifiers are diagrammed in Figs 2 and 3. Our results indicate that these phenotypes are complex polygenic traits affected by multiple alleles on multiple genes. In each decision tree, the notations displayed above each branch point indicate the loci and alleles used in the associated test criteria; the annotations below each leaf node indicate the most common serotype or host type, together with the actual numbers of isolates at the leaf having each phenotype. For example, in the serotype tree (Fig 2), the first branch point separates the isolates according to the allele at alignment position 9987. Those with an T allele are classified as serotype IE; the remainder are then tested at position 9201, with a C allele indicating serotype ID; the rest are tested at position 7764, with an A allele indicating serotype IAB and a G indicating serotype IC.

Depending on the true serotype of the isolate, serotype prediction accuracy ranged from 95.6% for ID to 98.5% for IAB and IE strains (Table 3). Serotype IE was almost universally associated with a T allele at position 9987, which is in the p6K/TF gene.

The SNP at position 9201 corresponds to residue 213 on the E2 protein; substituting G for C at this locus was shown previously to mediate a shift from serotype ID to IC [11]. Although we also found that this locus provided the best discrimination between serotypes ID and IAB/IC, the association was not as clear as indicated by the previous studies. Three serotype ID genomes (R16905, 8138, and 204381) had an A at position 9201, corresponding to a lysine at residue 213, which in earlier studies was associated with serotypes IAB and IC.

The SNP at position 7764 lies within the capsid gene; an A or G at this position, corresponding to lysine or arginine at residue 68, is associated with serotypes IAB and IC respectively. The only strains not classified correctly by this SNP are the three serotype ID

strains that are also misclassified by the SNP at 9201. The serotype data was obtained from previous studies, and it is possible that the serotypes were incorrectly determined.

To assess whether other loci would perform equally well for predicting serotype, we performed mutual information clustering to identify equivalence groups of loci, such that knowing the allele at one locus in a group completely determines the alleles of the other loci. A total of 4126 loci were present in the TC-83 genome and were polymorphic across all VEEV genomes. The largest equivalence group comprised 666 loci, which were those that have one allele for the serotype IE branch of the phylogenetic tree (including the 3 serotype ID outliers) and a different allele for the IAB/IC/ID branch. The remaining equivalence groups ranged in size from 2 to 76 loci; 2124 loci are singletons. The loci at positions 9987, 9201 and 7764 used in the serotype classifier are all singletons, having distinct patterns of alleles across the full set of isolates. The SNP at 9201 is also a singleton with respect to the 71 isolates in the non-IE subtree of the decision tree. However, across the 24 isolates in the IAB/IC subtree, there were 14 non-synonymous loci, 7 of which were in structural protein genes, which have the same allele pattern as 7764. Any of these loci would perform equally well in distinguishing serotype IAB and IC isolates, once the likely ID and IE isolates have been excluded by testing the loci at 9987 and 9201. Therefore, it would be premature to identify any one of these loci as determining the serotype IAB vs IC phenotype.

Host type prediction was less accurate than serotype prediction; 89.6% of strains were correctly predicted to have been collected from large mammals vs "small" hosts (mosquitos and rodents, including sentinel hamsters) (Table 3). This may reflect that hosts are sampled during outbreaks than during enzootic surveillance. The true positive rate (TPR) was larger for small hosts (95%) than for large hosts (81%). Close inspection of the SNP variants used in the decision

tree classifier (Fig 3) showed that their allele patterns were associated with phylogenetic branches rather than host type, and no mutations that universally associated with host type across multiple different phylogenetic branches could be identified.

We also built classifiers in which the predictors were restricted to non-synonymous loci within the genes encoding structural proteins, or further restricted to envelope protein genes. Serotype classifiers based on structural protein loci were more accurate than envelope glycoprotein-restricted classifiers (data not shown), but not as accurate as unrestricted classifiers. For host type prediction, the best overall classifier used envelope protein loci only, so restricting to smaller locus sets had no effect.

Table 3. Accuracy, positive (PPV) and negative predictive value (NPV), true positive (TPR) and negative (TNR) rates for serotype and host type predictions.

Phenotype	Accuracy	PPV	NPV	TPR	TNR
Serotype	94.8%	-	-	-	-
IAB	98.5%	90.0%	99.2%	90.0%	99.2%
IC	97.0%	71.4%	100.0%	100.0%	96.8%
ID	95.6%	97.9%	94.3%	90.2%	98.8%
IE	98.5%	98.4%	98.6%	98.4%	98.6%
Host type	87.4%	-	-	-	-
large	87.4%	90.7%	85.9%	75.0%	95.2%
small	87.4%	85.9%	90.7%	95.2%	75.0%

(PPV) = TP/(TP + FP), (NPV) = TN/(TN+FN), TPR=TP/(TP+FN), and TNR=TN/(TN+FP).

Comparison of single gene, MSA and SNP-based trees

We hypothesized that phylogenetic analyses of VEEV based on comparing single gene sequences, as was done in some earlier studies (2), would yield trees with lower resolution and differing topology than whole-genome MSA and SNP-based trees. To assess the impact of a single-gene approach, we compared the maximum parsimony tree based on all SNPs against trees generated using only the SNPs in each of the structural protein genes. We found that only 47% to 58% of the splits from the all-SNPs tree are present in any of the individual envelope gene trees (Table 4). Since only 9.5% to 14% of the SNPs occur within any of the envelope genes, the lower resolution of these trees is expected. The tree based on capsid gene SNPs had substantially worse resolution, however, with only 37% of the splits observed in the all-SNPs tree. Since the capsid gene contains over 3.5 times as many SNPs as the E3 gene, the number of splits shared by a gene-specific tree clearly depends on factors other than the total number of SNPs. The E1 gene resulted in the best representation of the tree, as it captures 58% of the splits identified in the all-SNPs tree.

Table 4. Comparison of trees from multiple sequence alignment versus all SNPs, and trees from SNPs located in a single gene versus all SNPs.

Tree comparison	Splits Found in 2nd tree	Total Splits in SNP tree	Fraction splits in SNP tree found in 2nd tree
All SNPs vs MSA	112	146	0.77
All SNPs vs E1	84	146	0.58
All SNPs vs E2	72	146	0.49
All SNPs vs E3	68	146	0.47
All SNPs vs capsid	54	146	0.37

To compare the topologies of trees generated with whole-genome SNPs or MSA to single-gene trees, we generated tanglegrams. S1 Fig shows a tanglegram with the MSA-based tree on the left and the all-SNPs tree on the right, with lines connecting the same taxa between trees. Differences between these trees were minor and within a reasonable expectation of uncertainty in the trees, mostly involving poorly resolved isolates such as Mucambo, CabassouCaAr, and PixunaBeAn. These isolates were collected from mosquito pools from 1954-1980 in French Guiana, Brazil, Argentina, and Peru, and are now considered different species in the VEE antigenic complex [26]. Each of these genomes has about 500 genome specific SNP alleles. They are the sole representatives of serotypes IF, IIIA, IIIB, IIIC, IV, V, and VI, each branching off the tree basal to the branches leading to the more heavily sequenced VEEV serotypes from Mexico, Peru, and Venezuela. In summary, the similarity between the whole genome SNP and MSA trees supports the SNP genotyping approach to phylogenetically characterize unsequenced samples using SNP arrays.

S2 and S3 Figs show tanglegrams with the all-SNPs tree on the left and the trees based on SNPs in the E1 (S2 Fig) or capsid (S3 Fig) gene on the right. The EEEV genomes were not clustered as a monophyletic group in any of the SNP gene trees, possibly because these genomes are too divergent from the VEEV genomes. Further, the capsid gene SNP tree had lower accuracy than the E1 gene tree, as indicated by the many crossing lines of the tanglegram in S3 Fig. The differences between the single gene and whole genome SNP trees illustrate the difficulty of phylogenetic analyses based on a small region rather than the full length of the genome, and suggest that SNP phylogenies based on single genes may have low resolution and accuracy.

Microarray analysis of VEEV cDNA samples

To address the question of whether microarrays provide a viable alternative to whole-genome sequencing for VEEV strain characterization, we developed a VEEV SNP array. The array included 70,760 probes covering all 7,926 loci discovered with kSNP. We hybridized cDNAs from 134 isolates to SNP arrays. Genome sequences were available for 81 of the isolates. We calculated overall concordance rates between the allele calls made by SNP microarray versus those called by whole genome sequences; these are summarized in S3 Table. The overall concordance rate was 96.2%. Hybridizations of replicate cDNA samples extracted from three isolates showed close agreement between replicates. The array correctly classified 76 out of 84 cDNA samples. Four of the 8 misclassified cases were highly similar sequences collected in the same location. One source of error was that the array analysis currently is not able to call a locus as missing, even if that locus is not present in the genome sequence, causing discordance between the genome and array genotypes.

A potential advantage of microarray analysis over DNA sequencing is its reduced need for viral isolation and culturing, allowing viruses to be characterized directly from a tissue sample. To test whether this was feasible, we isolated RNA from the brains of 3 replicate mice infected with VEEV strain TC-83, analyzed the RNA using the SNP microarray, and compared the array genotypes to our panel of 144 sequence-based genotypes. For all 3 replicates, the array genotypes were closest to those of the published sequence for the TC-83 strain, as shown in S4 Table. This suggests that a SNP microarray can produce accurate VEEV genotypes, even in the presence of a complex host DNA background.

Discussion

Tools for rapid genotyping of equine encephalitis virus strains and elucidating their phylogenetic relationships are critically important to understand why certain strains are likely to cause epizootic infection, and to forecast the incidence of potential epidemic events. The results above represent analyses of VEE complex strains derived from a wide range of hosts and geographic regions. The collected data indicate that our microarray and sequencing-based genotyping tools effectively distinguish VEEV strains and allow us to cluster those strains according to their derivation and phenotypic history.

Since the VEEV genome is small, whole genome multiple sequence alignment (MSA) of more than 140 sequences was feasible. Predicted genotype/phenotype associations were slightly more accurate when genotypes were based on variable positions in a whole genome alignment than when they were based on *k*-mer contexts defined by kSNP (data not shown). The MSA approach is usually not feasible for bacterial genomes, so that kSNP is typically a better option for bacterial genotype/phenotype association studies.

Relying on non-random associations between serotype and sequence variation, we were able to build decision trees to predict VEEV serotype. With 3 loci, prediction accuracy was 95.6% to 98.5%. However, strains that clustered phylogenetically with a different serotype were sometimes mislabeled by the decision trees. In addition, there were multiple loci that distinguished equally well between IAB and IC strains, after excluding isolates with a T at position 9987 (which are mostly serotype IE) or a C at position 9201 (which are mostly ID). These observations suggest that the actual amino acid variants that determine serotype may be any of a wide range of candidates, as suggested earlier [12], and that the association we observe between serotype and certain other variants is due to their co-inheritance with the causal variants.

We also noted that no variants associated perfectly with any of the serotypes; thus it must be possible to obtain the same shift in antigenic specificity from mutations at multiple loci.

A previous study [11] based on a smaller set of VEEV genomes investigated the mutations required for the virus to transition from the enzootic cycle (small mammals, *Culex* mosquitos, forest habitats) to the epizootic cycle (*Aedes/Psorophora* mosquitos, amplification in equids, transmission to humans). It reported that a single mutation in the E2 protein (T213 -> K or R), when engineered into a serotype ID enzootic strain, changed its serotype to IC and rendered it capable of causing enhanced viremia in horses, as well as possibly more efficiently infecting *Aedes* (*Ochlerotatus*) mosquitos implicated as vectors in equine-amplified epizootics. Our results, based on the larger set of VEEV genomes available now, suggest a more complex association between genotype and phenotype. We identified three serotype ID strains with the K213 E2 allele. There was no single locus that distinguished all of these ID strains from the IAB and IC strains nearest to them phylogenetically. This is further evidence that a variety of mutations can mediate the shift from ID to IAB or IC serotypes.

Comparison of the phylogenetic tree predicted from whole genome SNPs was similar to that from whole genome multiple sequence alignment. Narrowing to single gene SNP trees showed that the E1 gene SNPs more closely represent the whole genome SNP tree than do the SNPs from the other envelope protein or capsid genes. This concurs with previous analyses based on sequence alignment rather than SNPs, which also showed that the E1 gene captures the same high level relationships as the whole genome alignment but does not provide the same resolution [28]. However, these results emphasize that use of a small region of the genome for SNP analysis provides lower resolution than whole genome SNPs, and with some genes even results in different tree topology. A whole genome SNP approach more effectively represents complete phylogenetic relationships to reveal distinctions that would otherwise be overlooked.

A *k*-mer based approach to SNP discovery has limitations relative to full sequence alignment, particularly for highly variable RNA viruses. However, our comparison of data derived from multiple sequence alignments versus SNP analysis revealed that the resultant trees were very similar and reliably identified comparable splits. These observed similarities are important in that they support the use of our unique SNP array as an effective detection and genotyping tool without available whole genome sequence data. The SNP array results can be obtained within 24 hours as compared to 48-72 hours by whole genome sequencing. The cost of running a sample on SNP array is roughly 10 times less than whole genome sequencing. We have shown here that data obtained from SNP arrays are capable of reliably clustering strains in accordance with their respective whole genome sequence data. Array data provide sufficient accuracy in phylogenetic classification to correctly cluster isolates by clade and to identify the closest neighbors that have been sequenced or hybridized to the array. This technology would be particularly useful for rapidly evaluating a novel strain from an epizootic outbreak event. Further evaluation of the SNP array, using unknown or unsequenced VEEV strains, could provide additional validity and value of this technology in detection and genotyping of outbreak strains.

Acknowledgements

We thank Barry Hall for sharing the pre-release PPFS2 software.

References

1. Weaver SC, Ferro C, Barrera R, Boshell J, Navarro JC. Venezuelan equine encephalitis. Annu Rev Entomol. 2004;49:141-74. PubMed PMID: 14651460.

2. Quiroz E, Aguilar PV, Cisneros J, Tesh RB, Weaver SC. Venezuelan equine encephalitis in Panama: fatal endemic disease and genetic diversity of etiologic viral strains. PLoS Negl Trop Dis. 2009;3(6):e472. PubMed PMID: 19564908.

3. Vilcarromero S, Aguilar PV, Halsey ES, Laguna-Torres VA, Razuri H, Perez J, et al. Venezuelan equine encephalitis and 2 human deaths. Peru Emerg Infect Dis 2010;16:553-6.

4. Johnson KM, Shelokov A, Peralta PH, Dammin GJ, Young NA. Recovery of Venezuelan equine encephalomyelitis virus in Panama. A fatal case in man. The American journal of tropical medicine and hygiene. 1968;17(3):432-40. PubMed PMID: 5690051.

5. Ferro C, Olano VA, Ahumada M, Weaver S. [Mosquitos (Diptera: Culicidae) in the small village where a human case of Venezuelan equine encephalitis was recorded]. Biomedica. 2008;28(2):234-44. Epub 2008/08/23. doi: S0120-41572008000200008 [pii]. PubMed PMID: 18719725.

6. Aguilar PV, Greene IP, Coffey LL, Medina G, Moncayo AC, Anishchenko M, et al. Endemic Venezuelan Equine Encephalitis in Northern Peru. Emerging Infectious Diseases. 2004;10(5):880-8. doi: 10.3201/eid1005.030634. PubMed PMID: PMC3323213.

7. Forshey BM, Guevara C, Laguna-Torres VA, Cespedes M, Vargas J, Gianella A, et al. Arboviral Etiologies of Acute Febrile Illnesses in Western South America, 2000–2007. PLoS Negl Trop Dis. 2010;4(8):e787. doi: 10.1371/journal.pntd.0000787.

8. Vilcarromero S, Laguna-Torres A, Fernández C, Gotuzzo E, Suárez L, Céspedes M, et al. Venezuelan Equine Encephalitis and Upper Gastrointestinal Bleeding in Child. Emerging Infectious Diseases. 2009;15(2):323-5. doi: 10.3201/eid1502.081018. PubMed PMID: PMC2657634.

9. Watts DM, Lavera V, Callahan J, Rossi C, Oberste MS, Roehrig JT, et al. Venezuelan equine encephalitis and Oropouche virus infections among Peruvian army troops in the Amazon region of Peru. Am J Trop Med Hyg. 1997;56:661-7.

10. Watts DM, Lavera V, Callahan J, Rossi C, Oberste MS, Roehrig JT, et al. Venezuelan equine encephalitis febrile cases among humans in the Peruvian Amazon River region. . Am J Trop Med Hyg 1998;58:35-40.

11. Anishchenko M, Bowen RA, Paessler S, Austgen L, Greene IP, Weaver SC. Venezuelan encephalitis emergence mediated by a phylogenetically predicted viral mutation. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(13):4994-9. doi: 10.1073/pnas.0509961103.

12. Brault AC, Powers AM, Holmes EC, Woelk CH, Weaver SC. Positively Charged Amino Acid Substitutions in the E2 Envelope Glycoprotein Are Associated with the Emergence of

Venezuelan Equine Encephalitis Virus. Journal of Virology. 2002;76(4):1718-30. doi: 10.1128/JVI.76.4.1718-1730.2002. PubMed PMID: PMC135911.

13. Gardner S, Slezak T. Scalable SNP Analyses of 100+ Bacterial or Viral Genomes. J Forensic Res. 2010;1:107, doi:10.4172/2157-7145.1000107.

14. Gardner SN, Hall BG. When Whole-Genome Alignments Just Won't Work: kSNP v2 Software for Alignment-Free SNP Discovery and Phylogenetics of Hundreds of Microbial Genomes. PLoS ONE. 2013;8(12):e81760. doi: 10.1371/journal.pone.0081760.

15. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792-7.

16. Price MN. Fast Tree-Comparison Tools Berkeley, CA. Available from: http://meta.microbesonline.org/fasttree/treecmp.html.

17. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61(6):1061-7. doi: 10.1093/sysbio/sys062. PubMed PMID: 22780991.

18. Venkatachalam B, Apple J, St John K, Gusfield D. Untangling tanglegrams: comparing trees by their drawings. IEEE/ACM Trans Comput Biol Bioinform. 2010;7:588-97.

19. Gardner SN, Thissen J, McLoughlin K, Slezak T, Jaing C. Optimizing SNP microarray probe design for high accuracy microbial genotyping. J Microbio Meth. 2013:http://dx.doi.org/10.1016/j.mimet.2013.07.006.

20. Jaing C, Gardner SN, McLoughlin K, Mulakken N, Alegria-Hartman M, Banda P, et al. A functional gene array for detection of bacterial virulence elements. PLoS ONE. 2008;3(5):e2163. doi:10.1371/journal.pone.0002163.

21. Berge TO, Banks IS, Tigertt WD. Attenuation of Venezuelan equine encephalomyelitis virus by in vitro cultivation in guinea pig heart cells. Am J Hyg. 1961;73:209-18.

22. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria2014. Available from: http://www.R-project.org/.

23. Therneau T, Atkinson B, Ripley B, Oksanen J, De'ath G. mvpart: Multivariate partitioning. R package version 1.6-1 2013. Available from: http://cran.R-project.org/package=mvpart.

24. Venables W, Ripley B. Modern Applied Statistics with S. 4 ed. New York: Springer; 2002.

25. Kolaczkowski B, Thornton JW. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 2004;431:980-4.

26. Weaver SC, Barrett AD. Transmission cycles, host range, evolution and emergence of arboviral disease. Nature reviews Microbiology. 2004;2(10):789-801. doi: 10.1038/nrmicro1006. PubMed PMID: 15378043.

27. Weaver S, Pfeffer M, Marriott K, Kang W, Kinney R. Genetic evidence for the origins of Venezuelan equine encephalitis virus subtype IAB outbreaks. Am J Trop Med Hyg. 1999;60(3):441-8.

28. Wolfe DN, Heppner DG, Gardner SN, Jaing C, Dupuy LC, Schmaljohn CS, et al. Current Strategic Thinking for the Development of a Trivalent Alphavirus Vaccine for Human Use. The American Journal of Tropical Medicine and Hygiene. 2014. doi: 10.4269/ajtmh.14-0055.

Figure captions

Fig 1. SNP phylogeny of VEEV isolates by parsimony. Strains are labeled by serotype-country-year collected-strain-host. Country codes are GA=Guatemala, PE=Peru, NI=Nicaragua, VE=Venezuela, CO=Colombia, TR=Trinidad, PA=Panama, US=USA, EC=Ecuador, ME=Mexico, BE=Belize, HO=Honduras, BR=Brazil, AR=Argentina, FG=French Guiana. Host codes are hor=horse, don=donkey, hum=human, mos=mosquito, ham=hamster, mus=mouse. u=unknown. Strains are colored by serotype (blue=IE, green=ID, red=IC, and purple=IAB). Hosts from which the strains were collected are indicated with symbols at the branch tips (red circles=human, orange circles= horses, blue circles=mosquitos, and green squares=hamsters). Counts of the number of alleles shared uniquely by the sequences down each branch are shown at the nodes in blue.

Fig 2. Decision tree for prediction of serotype from SNP alleles. Notations above internal nodes indicate SNP position in the TC-83 genome and alleles corresponding to left and right branches. Numbers below terminal nodes are numbers of isolates in node with serotypes IAB/IC/ID/IE respectively.

Fig 3. Decision tree for prediction of host type from SNP alleles. Notations above internal nodes are as in Fig 2. Numbers below terminal nodes are numbers of isolates in node collected from large/small host types, respectively.

Supporting Information Legends

S1 Fig. Tanglegram connecting the corresponding taxa which illustrates the high similarity between the MSA tree (left) and the SNP tree (right).

S2 Fig. Tanglegram illustrating where the SNP tree based on all the SNPs (left) and that based only on the SNPs in the E1 gene (right) differ.

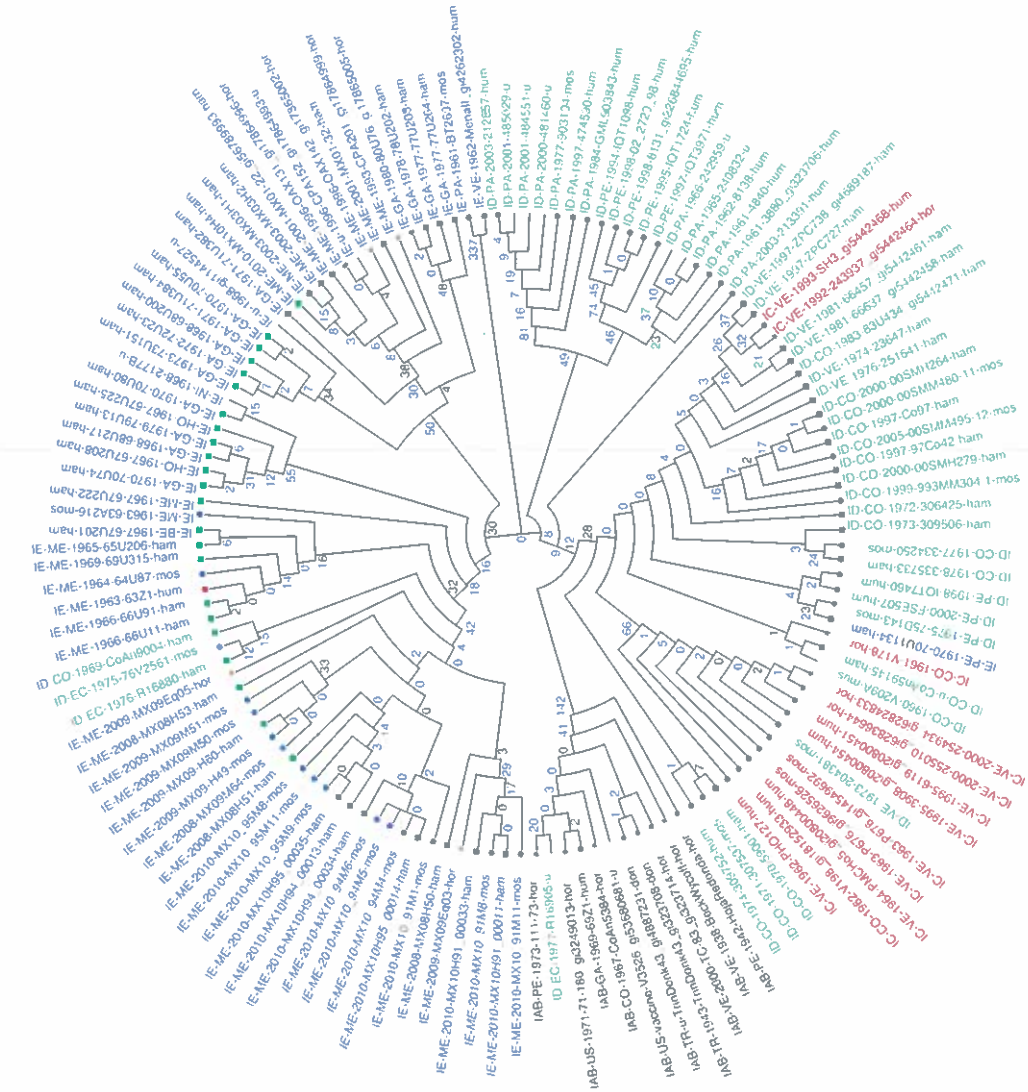
S3 Fig. Tanglegram illustrating where the SNP tree based on all the SNPs (left) and that based only on the SNPs in the capsid gene (right) differ.

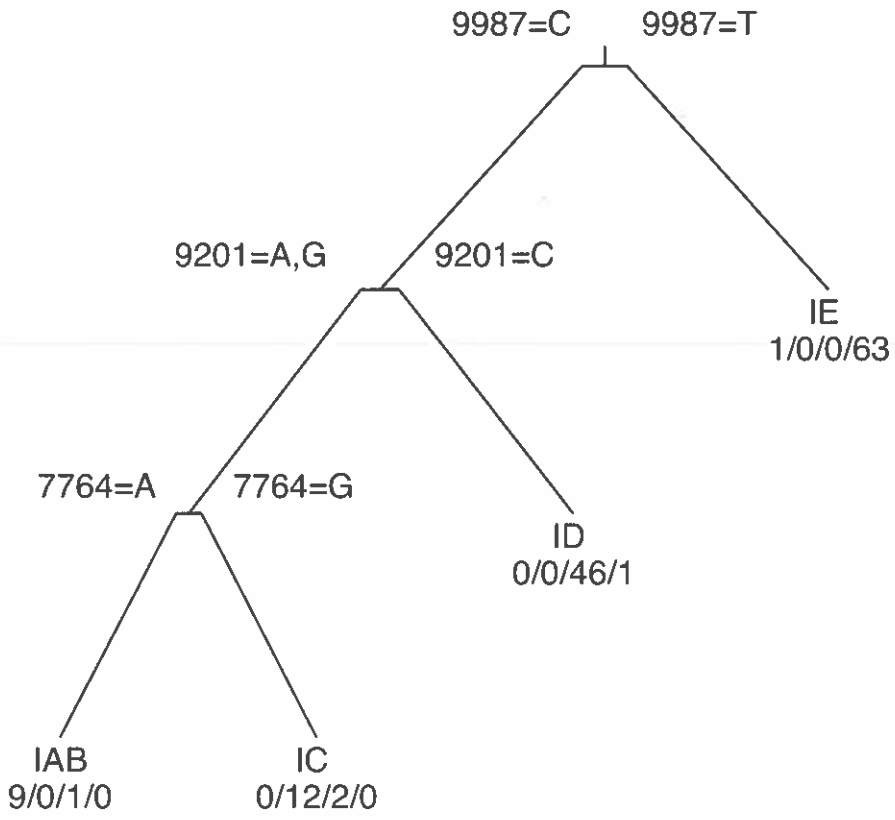
S1 Table. Characteristics of VEE antigenic complex strains used for whole genome SNP analysis and/or tested on SNP microarray.

S2 Table. Annotations, 13-mer contexts and reference genome alignments for SNPs identified by whole genome analysis.

S3 Table. Concordance of array and genome-based allele calls, for non-passaged isolates with known genome sequences. Rows in bold text indicate replicate arrays for the same isolates.

S4 Table. Comparison of genotypes for VEEV on tissue from TC-83 infected mice.




9987=C
9987=T
9201=A,G
9201=C
IE
1/0/0/63
7764=A
7764=G
ID
0/0/46/1
IAB
9/0/1/0
IC
0/12/2/0

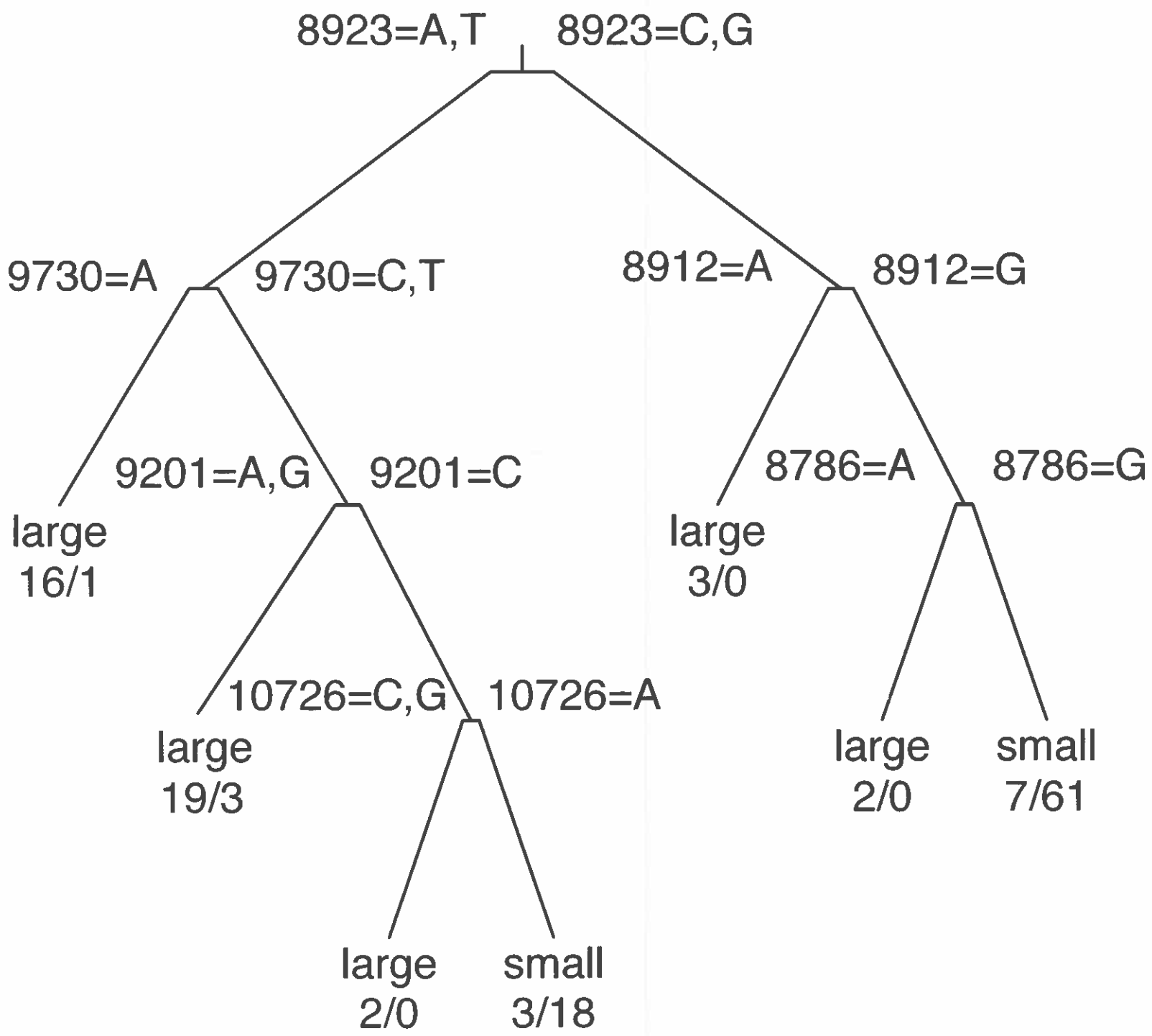

8923=A,T
8923=C,G
9730=A
9730=C,T
8912=A
8912=G
9201=A,G
9201=C
8786=A
8786=G
large
16/1
large
3/0
10726=C,G
10726=A
large
19/3
large
2/0
small
7/61
large
2/0
small
3/18

Figure S1

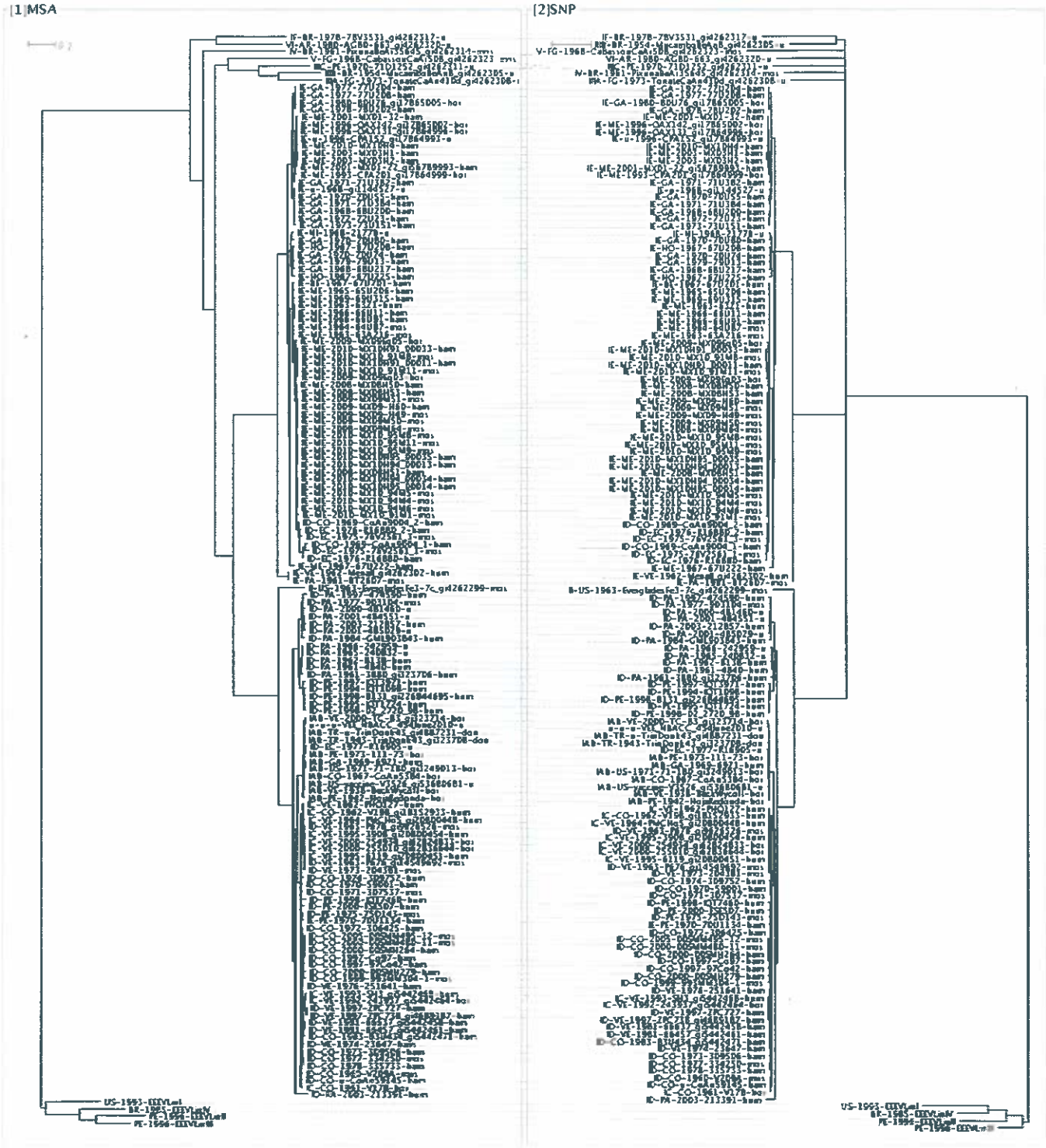


Figure S2

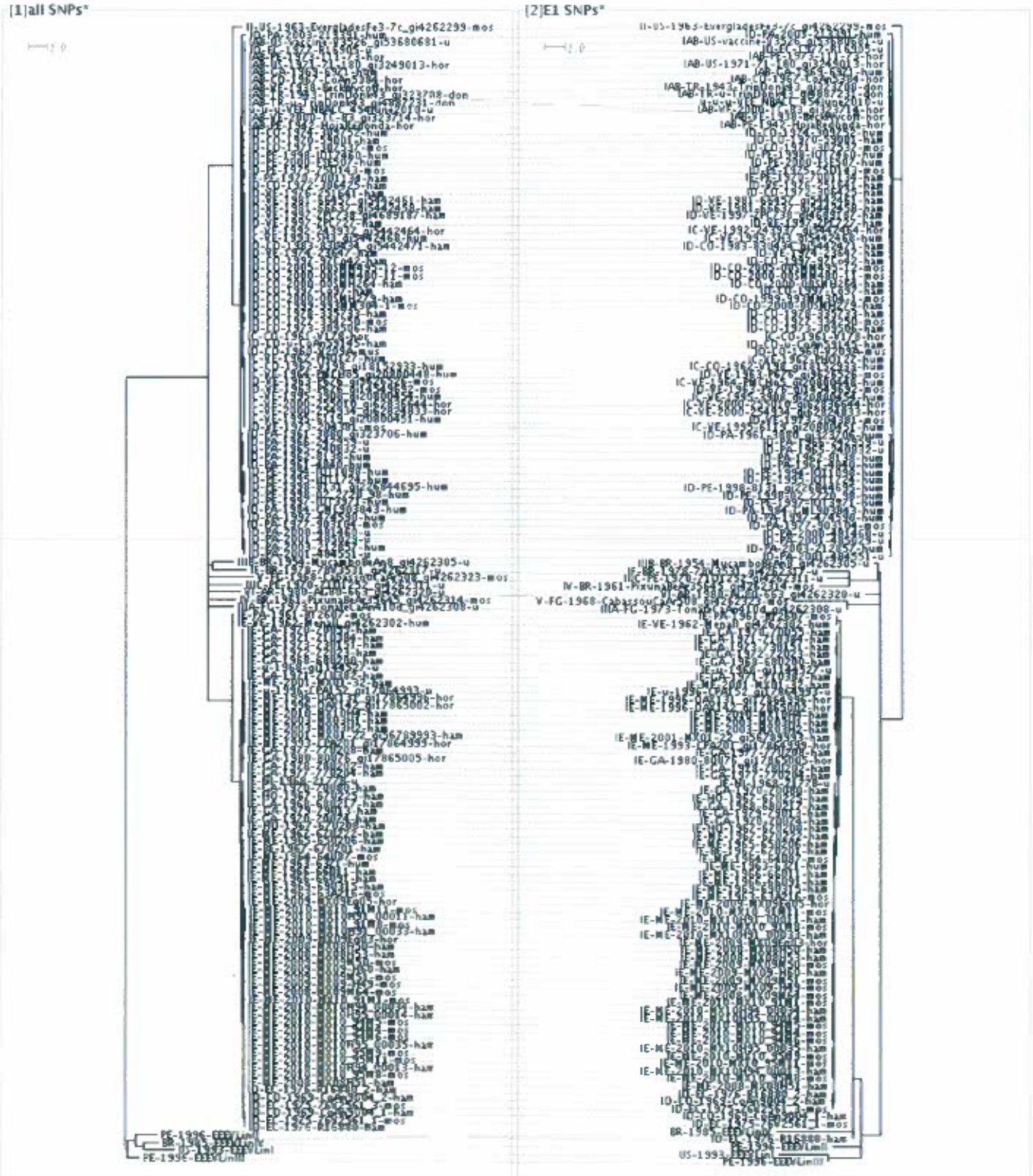
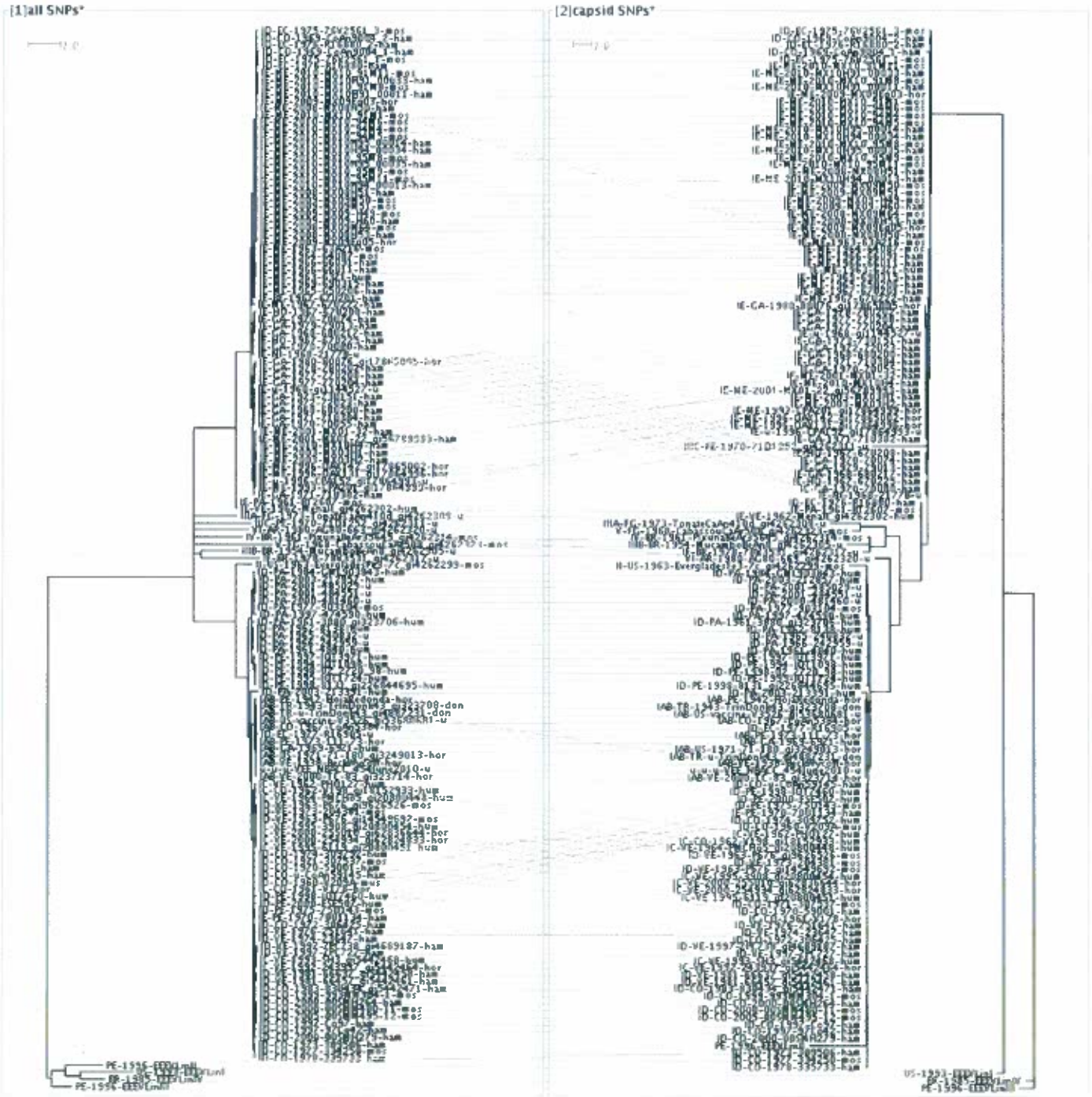


Figure S3



Serotype	Sequence ID	Accession	gi number	Passage history	Year of collection	Host	Location where collected	Country	Identifier in tree	Tested on SNP microarray	Genome available	Other data	Country key	Host key	Passage history key
IAB	CoAn5384	KC344485.1	449535945	sm2,cec1	1967	hor	Cali, Colombia	CO	IAB-CO-1967-CoAn5384-hor	YES	YES		AR=Argentina	don=donkey	CEC = chicken embryo cell culture
IAB	6971	KC344505.1	449536005	sm2,BHK1	1969	hum	Guatemala	GA	IAB-GA-1969-6971-hum	YES	YES		BE=Belize	ham=hamster	C6/36 = Aedes albopictus mosquito
IAB	111-73	KC344483.1	449535939	sm3	1973	hor	Peru	PE	IAB-PE-1973-111-73-hor	YES	YES		BR=Brazil	hor=horse	Sm (or SMB) = suckling mouse brain
IAB	Beck_Wycoff	KC344516.1	449536038	sm8, cec1	1938	hor	Aragua St., Venezuela	VE	IAB-VE-1938-BeckWycoff-hor	YES	YES		CO=Colombia	hum=human	U = unknown
IC	V178	KC344484.1	449535942	sm1, V1	1961	hor	Cundinamarca, Colombia	CO	IC-CO-1961-V178-hor	YES	YES		EC=Ecuador	mos=mosquito	V = Vero African green monkey kidney
IC	V198_gi18152933	U55342.2	18152933	cec7	1962	hum		CO	IC-CO-1962-V198_gi18152933-hum	YES	YES		FG=French Guiana	mule=mule	BHK = baby hamster kidney
IC	PHO127	KC344528.1	449536074	BHK1	1962	hum	Guajira, Venezuela	VE	IC-VE-1962-PHO127-hum	YES	YES		GA=Guatemala	mus=mouse	RK = rabbit kidney
IC	P676_gi14549692	AF375051.1	14549692	u	1963	mos		VE	IC-VE-1963-P676_gi14549692-mos	YES	YES	brain	HO=Honduras	rat=Proechimys sp	LLC1 = Lewis lung carcinoma line 1
IC	PMCHoS_gi20800448	U55345.2	20800448	u	1964	hum	Monagas, Venezuela	VE	IC-VE-1964-PMCHoS_gi20800448-hum	YES	YES		ME=Mexico	u=unknown	
IC	243937_gi5442464	AF004459.2	5442464	V1, BHK1	1992	hor	Trujillo State, Venezuela	VE	IC-VE-1992-243937_gi5442464-hor	YES	YES		NI=Nicaragua		
IC	SH3_gi5442468	U55360.2	5442468	V1	1993	hum	Candelaria, Venezuela	VE	IC-VE-1993-SH3_gi5442468-hum	YES	YES		PA=Panama		
IC	6119_gi20800451	U55347.2	20800451	V1	1995	hum	Falcon State, Venezuela	VE	IC-VE-1995-6119_gi20800451-hum	YES	YES		PE=Peru		
IC	254934_gi62824833	AY973943.1	62824833	BHK1	2000	hor	Barinas State, Venezuela	VE	IC-VE-2000-254934_gi62824833-hor	YES	YES		TR=Trinidad		
IC	255010_gi62836644	AY986475.1	62836644	sm2,V1	2000	hor	Barinas State, Venezuela	VE	IC-VE-2000-255010_gi62836644-hor	YES	YES		US=USA		
ID	V209A	AF004465.1	2564173	sm2,V2	1960	mus	u	CO	ID-CO-1960-V209A-mus	YES	YES		VE=Venezuela		
ID	CoAn9004_1	KC344530.1	449536080	sm3, V1	1969	ham	Tumaco, Columbia	CO	ID-CO-1969-CoAn9004_1-ham	YES	YES		u=unknown		
ID	307537	KC344519.1	449536047	V1	1971	mos	Pto. Boyaca, Colombia	CO	ID-CO-1971-307537-mos	YES	YES	Mansonia sp.			
ID	309506	KC344520.1	449536050	V1	1973	ham	Pto. Boyaca, Colombia	CO	ID-CO-1973-309506-ham	YES	YES				
ID	309752	KC344477.1	449535922	cec1	1974	hum	Lozania, Colombia	CO	ID-CO-1974-309752-hum	YES	YES				
ID	334250	KC344487.1	449535951	V2	1977	mos	Pto. Boyaca, Colombia	CO	ID-CO-1977-334250-mos	YES	YES	Ae. fulvus			
ID	335733	KC344514.1	449536032	none	1978	ham	Pto. Boyaca, Colombia	CO	ID-CO-1978-335733-ham	YES	YES				
ID	97Co42	KC344523.1	449536059	v1	1997	ham	Monte San Miguel, Colombia	CO	ID-CO-1997-97Co42-ham	YES	YES	heart			
ID	993MM304-1	KC344521.1	449536053	none	1999	mos	Monte San Miguel, Colombia	CO	ID-CO-1999-993MM304-1-mos	YES	YES				
ID	00SMH264	KC344459.1	449535868	none	2000	ham	Monte San Miguel, Colombia	CO	ID-CO-2000-00SMH264-ham	YES	YES	heart			
ID	CoAn59145	KC344524.1	449536062	BHK3	u	ham	Tibu, Colombia	CO	ID-CO-u-CoAn59145-ham	YES	YES				
ID	76V2561	KC344531.1	449536084	sm4	1975	mos	u	EC	ID-EC-1975-76V2561-mos	YES	YES				
ID	R16880	KC344529.1	449536077	sm4, V1	1976	ham	u	EC	ID-EC-1976-R16880-ham	YES	YES				
ID	3880_gi323706	L00930.1	323706	u	1961	hum	Canito, Panama	PA	ID-PA-1961-3880_gi323706-hum	YES	YES	fatal			
ID	4840	KC344506.1	449536008	BHK1,sm2,V1,cec1	1961	hum	PA	PA	ID-PA-1961-4840-hum	YES	YES				
ID	8138	KC344471.1	449535904	cec2	1962	hum	El Rincon, Panama	PA	ID-PA-1962-8138-hum	YES	YES				
ID	242959	KC344488.1	449535954	u	1966	u	Gamboa, Panama	PA	ID-PA-1966-242959-u	YES	YES				
ID	903104	KC344503.1	449535999	BHK1	1977	mos	Bayano, Panama	PA	ID-PA-1977-903104-mos	YES	YES	Cx. atkeni s.l.			
ID	GML903843	KC344472.1	449535907	V1, BHK1	1984	hum	Bayano, Panama	PA	ID-PA-1984-GML903843-hum	YES	YES				
ID	474590	KC344473.1	449535910	V2	1997	hum	P. metro, Panama	PA	ID-PA-1997-474590-hum	YES	YES	encephalatis			
ID	481460	KC344510.1	449536020	V2	2000	u	Peste, Panama	PA	ID-PA-2000-481460-u	YES	YES				
ID	484551	KC344511.1	449536023	V2	2001	u	Darien, Panama	PA	ID-PA-2001-484551-u	YES	YES				
ID	485029	KC344474.1	449535913	V2	2001	u	Darien, Panama	PA	ID-PA-2001-485029-u	YES	YES				
ID	212857	KC344475.1	449535916	SMB-1	2003	hum	Darien, Panama	PA	ID-PA-2003-212857-hum	YES	YES				
ID	213391	KC344476.1	449535919	SMB-1	2003	hum	B del Toro, Panama	PA	ID-PA-2003-213391-hum	YES	YES	encephalatis			
ID	75D143	KC344525.1	449536065	cec1	1975	mos	Iquitos, Peru	PE	ID-PE-1975-75D143-mos	YES	YES				
ID	IQT1724	KC344490.1	449535960	V1	1995	hum	Loreto, Peru	PE	ID-PE-1995-IQT1724-hum	YES	YES				
ID	02_2720_98	KC344504.1	449536002	C6/36-1	1998	hum	Iquitos, Peru	PE	ID-PE-1998-02_2720_98-hum	YES	YES				
ID	FSE507	KC344522.1	449536056	V1	2000	hum	Iquitos, Peru	PE	ID-PE-2000-FSE507-hum	YES	YES				
ID	204381	KC344512.1	449536026	u	1973	mos	Delta Amacuro, Venezuela	VE	ID-VE-1973-204381-mos	YES	YES				
ID	23647	KC344508.1	449536014	V1, BHK1	1974	ham	Catatumbo, Venezuela	VE	ID-VE-1974-23647-ham	YES	YES				
ID	251641	KC344509.1	449536017	sm3,V3	1976	ham	Pto. Concha, Venezuela	VE	ID-VE-1976-251641-ham	YES	YES				
ID	ZPC727	KC344513.1	449536029	none	1997	ham	Las Nubes, catatumbo, venezue	VE	ID-VE-1997-ZPC727-ham	YES	YES				
IE	67U201	KC344433.1	449535792	sm1	1967	ham	Belize	BE	IE-BE-1967-67U201-ham	YES	YES				
IE	68U200	KC344435.1	449535797	none	1968	ham	La Avellana, Santa Rosa Departr	GA	IE-GA-1968-68U200-ham	YES	YES				
IE	68U217	KC344442.1	449535817	BHK1	1968	ham	Pto. Barrios, Guatemala	GA	IE-GA-1968-68U217-ham	YES	YES				
IE	70U74	KC344443.1	449535820	u	1970	ham	Pto. Barrios, Guatemala	GA	IE-GA-1970-70U74-ham	YES	YES				
IE	71U382	KC344466.1	449535889	V1	1971	ham	La Avellana, Guatemala	GA	IE-GA-1971-71U382-ham	YES	YES				
IE	71U384	KC344454.1	449535853	sm1, V1	1971	ham	Santa Rosa Department, Guater	GA	IE-GA-1971-71U384-ham	YES	YES				
IE	72U23	KC344467.1	449535892	V1	1972	ham	La Avellana, Guatemala	GA	IE-GA-1972-72U23-ham	YES	YES				
IE	73U151	KC344434.1	449535795	u	1973	ham	La Avellana, Guatemala	GA	IE-GA-1973-73U151-ham	YES	YES				
IE	78U202	KC344438.1	449535806	u	1978	ham	La Avellana, Guatemala	GA	IE-GA-1978-78U202-ham	YES	YES				
IE	79U13	KC344455.1	449535856	BHK 1	1979	ham	Izabal Department, Guatemala	GA	IE-GA-1979-79U13-ham	YES	YES				
IE	80U76_gi17865005	AF448539.1	17865005	u	1980	hor	La Avellana, Guatemala	GA	IE-GA-1980-80U76_gi17865005-hor	YES	YES				
IE	67U208	KC344456.1	449535859	V?	1967	ham	Honduras	HO	IE-HO-1967-67U208-ham	YES	YES				
IE	67U225	KC344444.1	449535823	sm1	1967	ham	Pto. Cortez, Honduras	HO	IE-HO-1967-67U225-ham	YES	YES				
IE	63A216	KC344457.1	449535862	sm1	1963	mos	Veracruz, Mexico	ME	IE-ME-1963-63A216-mos	YES	YES				
IE	63Z1	KC344431.1	449535786	sm1, V1	1963	hum	Veracruz, Mexico	ME	IE-ME-1963-63Z1-hum	YES	YES				
IE	65U206	KC344446.1	449535829	sm1	1965	ham	Sontecomapan, Mexico	ME	IE-ME-1965-65U206-ham	YES	YES				
IE	66U91	KC344447.1	449535832	cec1	1966	ham	Sontecomapan, Mexico	ME	IE-ME-1966-66U91-ham	YES	YES				
IE	67U222	KC344439.1	449535809	V1	1967	ham	Minatitlan, Mexico	ME	IE-ME-1967-67U222-ham	YES	YES				
IE	69U315	KC344448.1	449535835	sm1	1969	ham	Sontecomapan, Mexico	ME	IE-ME-1969-69U315-ham	YES	YES				
IE	MX01-32	KC344515.1	449536035	none	2001	ham	Chiapas State, Mexico	ME	IE-ME-2001-MX01-32-ham	YES	YES				
IE	MX03H1	KC344482.1	449535937	none	2003	ham	Las Coaches, Pijijiapan, Chiapas	ME	IE-ME-2003-MX03H1-ham	YES	YES				
IE	MX03H2	KC344464.1	449535883	none	2003	ham	Las Coaches, Pijijiapan, Chiapas	ME	IE-ME-2003-MX03H2-ham	YES	YES				
IE	MX08H50	KC344449.1	449535838	V1	2008	ham	E. Coachapa, Minatitlan, Veracr	ME	IE-ME-2008-MX08H50-ham	YES	YES				
IE	MX08H53	KC344451.1	449535844	V1	2008	ham	Tacoteno, Minatitlan, Veracruz	ME	IE-ME-2008-MX08H53-ham	YES	YES				
IE	MX09M64	KC344480.1	449535931	V1	2008	mos	Tacoteno, Minatitlan, Veracruz	ME	IE-ME-2008-MX09M64-mos	YES	YES	Cx. taeniopus			
IE	MX09Eq03	KC344481.1	449535934	V1	2009	hor	Tacoteno, Minatitlan, Veracruz	ME	IE-ME-2009-MX09Eq03-hor	YES	YES				
IE	MX10H91_00011	KC344468.1	449535895	none	2010	ham	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10H91_00011-ham	YES	YES	heart			
IE	2177B	KC344441.1	449535814	V3	1968	u	Nicaragua	NI	IE-NI-1968-2177B-u	YES	YES				
IE	BT2607	KC344432.1	449535789	u	1961	mos	Almirante, Panama	PA	IE-PA-1961-BT2607-mos	YES	YES	Cx. taeniopus			
IE	70U1134	KC344486.1	449535948	u	1970	ham	Iquitos, Peru	PE	IE-PE-1970-70U1134-ham	YES	YES				
IE	Menali_gi4262302	AF075252.1	4262302	u	1962	hum	Zulia State, Venezuela	VE	IE-VE-1962-Menali_gi4262302-hum	YES	YES				
IIIB	MucamboBeAn8_gi42623	AF075253.1	4262305	u	1954	u	Brazil	BR	IIIB-BR-1954-MucamboBeAn8_gi4262305-u	YES	YES				
IV	PixunaBeAr35645_gi4262	AF075256.1	4262314	sm4, V1	1961	mos	Brazil	BR	IV-BR-1961-PixunaBeAr35645_gi4262314-m	YES	YES	Anopheles nimbus			
V	CabassouCaAr508_gi4262	AF075259.1	4262323	sm10	1968	mos	French Guiana (Cabassou)	FG	V-FG-1968-CabassouCaAr508_gi4262323-mo	YES	YES	Culex portesi			
IAB	HojaRedonda	KC344430.1	449535783	u	1942	hor	Hoja Redonda, Ica, Peru	PE	IAB-PE-1942-HojaRedonda-hor	NO	YES				
IAB	TrinDonk43_gi323708	J04332.1	323708	u	1943	don	Trinidad	TR	IAB-TR-1943-TrinDonk43_gi323708-don	NO	YES				
IAB	TrinDonk43_gi4887231	L01442.2	4887231	u	1943	don	Trinidad	TR	IAB-TR-u-TrinDonk43_gi4887231-don	NO	YES				
IAB	71-180_gi3249013	AF069903.1	3249013	u	1971	hor	Texas	US	IAB-US-1971-71-180_gi3249013-hor	NO	YES				
IAB	V3526_gi53680681	AY741139.1	53680681	u	vaccine	u	US	US	IAB-US-vaccine-V3526_gi53680681-u	NO	YES	vaccine			
IAB	TC-83_gi323714	L01443.1	323714	sm2	1977	hor	Carabobo State, Venezuela	VE	IAB-VE-1977-TC-83_gi323714-hor	NO	YES	vaccine			
IC	P676_gi9626526	NC_001449.1	9626526	u	1963	mos	Miranda State, Venezuela	VE	IC-VE-1963-P676_gi9626526-mos	NO	YES				
IC	3908_gi20800454	U55350.2	20800454	V1	1995	hum	Zulia State, Venezuela	VE	IC-VE-1995-3908_gi20800454-hum	NO	YES				

ID	59001	AF004468.1	2564179	V1	1970	ham	Pto. Boyaca, Colombia	CO	ID-CO-1970-59001-ham	NO	YES	
ID	306425	KC344502.1	339535996	V2	1972	ham	Pto. Boyaca, Colombia	CO	ID-CO-1972-306425-ham	NO	YES	
ID	83U434_gi5442471	U55362.2	5442471	cec1,BHK1	1983	ham	Rio de Oro, Colombia	CO	ID-CO-1983-83U434_gi5442471-ham	NO	YES	
ID	Co97	KC344429.1	339535780	V1	1997	ham	Monte San Miguel, Colombia	CO	ID-CO-1997-Co97-ham	NO	YES	
ID	005MH279	KC344461.1	339535874	none	2000	ham	Monte San Miguel, Colombia	CO	ID-CO-2000-005MH279-ham	NO	YES	heart
ID	005MM480-11	KC344460.1	339535871	none	2000	mos	Monte San Miguel, Colombia	CO	ID-CO-2000-005MM480-11-mos	NO	YES	
ID	005MM495-12	KC344462.1	339535877	none	2005	mos	Monte San Miguel, Colombia	CO	ID-CO-2005-005MM495-12-mos	NO	YES	
ID	R16905	KC344517.1	339536041	sm5	1977	u	EC	EC	ID-EC-1977-R16905-u	NO	YES	
ID	240832	KC344518.1	339536044	u	1965	u	Gamboa, Panama	PA	ID-PA-1965-240832-u	NO	YES	
ID	IQT1098	KC344526.1	339536068	sm1, LtC1	1994	hum	Iquitos, Peru	PE	ID-PE-1994-IQT1098-hum	NO	YES	
ID	IQT3971	KC344507.1	339536011	C6/36-1	1997	hum	Iquitos, Peru	PE	ID-PE-1997-IQT3971-hum	NO	YES	
ID	8131_gi226844695	DQ390224.2	226844695	u	1998	hum	Belen, Iquitos, Peru	PE	ID-PE-1998-8131_gi226844695-hum	NO	YES	
ID	IQT7460	AY966910.2	111116767	u	1998	hum	u	PE	ID-PE-1998-IQT7460-hum	NO	YES	
ID	66457_gi5442461	AF004472.2	5442461	V1	1981	ham	Zulia State, Venezuela	VE	ID-VE-1981-66457_gi5442461-ham	NO	YES	
ID	66637_gi5442458	AF004458.2	5442458	V1	1981	ham	Zulia State, Venezuela	VE	ID-VE-1981-66637_gi5442458-ham	NO	YES	
ID	ZPC738_gi4689187	AF100566.1	4689187	V1	1997	ham	Zulia State, Venezuela	VE	ID-VE-1997-ZPC738_gi4689187-ham	NO	YES	
IE	70U55	KC344436.1	339535800	sm1	1970	ham	La Avellana, Guatemala	GA	IE-GA-1970-70U55-ham	NO	YES	
IE	70U80	KC344458.1	339535865	sm1	1970	ham	Izabal Department, Guatemala	GA	IE-GA-1970-70U80-ham	NO	YES	
IE	77U204	KC344527.1	339536071	BHK1	1977	ham	La Avellana, Guatemala	GA	IE-GA-1977-77U204-ham	NO	YES	
IE	77U208	KC344437.1	339535803	u	1977	ham	La Avellana, Guatemala	GA	IE-GA-1977-77U208-ham	NO	YES	
IE	64U87	KC344445.1	339535826	sm1	1964	mos	Sontecomapan, Mexico	ME	IE-ME-1964-64U87-mos	NO	YES	Cx. opisthopus
IE	66U11	KC344440.1	339535811	cec1	1966	ham	Minatitlan, Mexico	ME	IE-ME-1966-66U11-ham	NO	YES	
IE	CPA201_gi17864999	AF448537.1	17864999	RK1,sm1	1993	hor	Rancho El Recuerdo, Mapastepe	ME	IE-ME-1993-CPA201_gi17864999-hor	NO	YES	
IE	OAX131_gi17864996	AF448536.1	17864996	sm1,RK1	1996	hor	Oaxaca State, Mexico	ME	IE-ME-1996-OAX131_gi17864996-hor	NO	YES	
IE	OAX142_gi17865002	AF448538.1	17865002	sm1,RK1	1996	hor	Tapanatepec, Oaxaca, Mexico	ME	IE-ME-1996-OAX142_gi17865002-hor	NO	YES	brain
IE	MX01-22_gi56789993	AY823299.1	56789993	none	2001	ham	Chiapas State, Mexico	ME	IE-ME-2001-MX01-22_gi56789993-ham	NO	YES	
IE	MX08H51	KC344450.1	339535841	V1	2008	ham	E. Coachapa, Minatitlan, Veracr	ME	IE-ME-2008-MX08H51-ham	NO	YES	
IE	MX09-H49	KC344453.1	339535850	V1	2009	mos	Tacoteno, Minatitlan, Veracruz	ME	IE-ME-2009-MX09-H49-mos	NO	YES	Cq. nigricans
IE	MX09-H60	KC344452.1	339535847	V1	2009	ham	Tacoteno, Minatitlan, Veracruz	ME	IE-ME-2009-MX09-H60-ham	NO	YES	
IE	MX09Eq05	KC344463.1	339535880	V1	2009	hor	Tacoteno, Minatitlan, Veracruz	ME	IE-ME-2009-MX09Eq05-hor	NO	YES	
IE	MX09M50	KC344479.1	339535928	V1	2009	mos	Tacoteno, Minatitlan, Veracruz	ME	IE-ME-2009-MX09M50-mos	NO	YES	M. trivittatus
IE	MX09M51	KC344478.1	339535925	none	2009	mos	Tacoteno, Minatitlan, Veracruz	ME	IE-ME-2009-MX09M51-mos	NO	YES	Cx. nigripalus
IE	MX10_91M1	KC344493.1	339535969	none	2010	mos	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10_91M1-mos	NO	YES	pool
IE	MX10_91M11	KC344495.1	339535975	none	2010	mos	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10_91M11-mos	NO	YES	pool
IE	MX10_91M8	KC344494.1	339535972	none	2010	mos	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10_91M8-mos	NO	YES	pool
IE	MX10_94M4	KC344496.1	339535978	none	2010	mos	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10_94M4-mos	NO	YES	pool
IE	MX10_94M5	KC344497.1	339535981	none	2010	mos	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10_94M5-mos	NO	YES	pool
IE	MX10_94M6	KC344498.1	339535984	none	2010	mos	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10_94M6-mos	NO	YES	pool
IE	MX10_95M11	KC344501.1	339535993	none	2010	mos	u	ME	IE-ME-2010-MX10_95M11-mos	NO	YES	
IE	MX10_95M8	KC344499.1	339535987	none	2010	mos	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10_95M8-mos	NO	YES	pool
IE	MX10_95M9	KC344500.1	339535990	none	2010	mos	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10_95M9-mos	NO	YES	pool
IE	MX10H4	KC344465.1	339535886	none	2010	ham	El Dorado, Mapastepec, Chiapa	ME	IE-ME-2010-MX10H4-ham	NO	YES	
IE	MX10H91_00033	KC344468.1	339535895	none	2010	ham	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10H91_00033-ham	NO	YES	heart
IE	MX10H94_00013	KC344491.1	339535963	none	2010	ham	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10H94_00013-ham	NO	YES	heart
IE	MX10H94_00034	KC344469.1	339535898	V1	2010	ham	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10H94_00034-ham	NO	YES	heart
IE	MX10H95_00014	KC344470.1	339535901	none	2010	ham	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10H95_00014-ham	NO	YES	heart
IE	MX10H95_00035	KC344470.1	339535901	none	2010	ham	Minatitlan, Veracruz State, Mex	ME	IE-ME-2010-MX10H95_00035-ham	NO	YES	heart
IE	gi1144527	U34999.1	1144527	u	1968	u	u	u	IE-u-1968-gi1144527-u	NO	YES	
IE	CPA152_gi17864993	AF448535.1	17864993	u	1996	u	u	u	IE-u-1996-CPA152_gi17864993-u	NO	YES	
IF	78V3531_gi4262317	AF075257.1	4262317	u	1978	u	Brazil	BR	IF-BR-1978-78V3531_gi4262317-u	NO	YES	
II	EvergladesFe3-7c_gi4262	AF075251.1	4262299	u	1963	mos	Florida, USA	US	II-US-1963-EvergladesFe3-7c_gi4262299-mo	NO	YES	(Culex (Melanoconion) sp.)
IIIA	TonateCaAn410d_gi4262	AF075254.1	4262308	u	1973	u	French Guiana	FG	IIIA-FG-1973-TonateCaAn410d_gi4262308-u	NO	YES	
IIIC	71D1252_gi4262311	AF075255.1	4262311	u	1970	u	PE	PE	IIIC-PE-1970-71D1252_gi4262311-u	NO	YES	
u	VEE_NBACC_454June2010	unpublished	unpublished	u	u	u	u	u	u-u-u-VEE_NBACC_454June2010-u	NO	YES	
VI	AG80-663_gi4262320	AF075258.1	4262320	u	1980	u	Argentina	AR	VI-AR-1980-AG80-663_gi4262320-u	NO	YES	
IAB	Piura			sm3	1942	mule	Piura, Peru	PE	IAB-PE-1942-Piura-mule	YES	NO	
IAB	TRD			sm7	1943	don	Trinidad	TR	IAB-TR-1943-TRD-don	YES	NO	
IAB	V-263E			u	1943	don	Trinidad	TR	IAB-TR-1943-V-263E-don	YES	NO	
IAB	E541_73			sm1, cec2	1973	hum	Guajira, Zulia State, Venezuela	VE	IAB-VE-1973-E541_73-hum	YES	NO	
IC	V202			sm1,V1	1962	hum	Guajira, Colombia	CO	IC-CO-1962-V202-hum	YES	NO	
IC	369673			V1	1999	hum	Manaure, Guajira, Colombia	CO	IC-CO-1999-369673-hum	YES	NO	
IC	369676			V1	1999	hum	Manaure, Guajira, Colombia	CO	IC-CO-1999-369676-hum	YES	NO	
IC	369678			V1	1999	hum	Manaure, Guajira, Colombia	CO	IC-CO-1999-369678-hum	YES	NO	
IC	369680			V1	1999	hum	Manaure, Guajira, Colombia	CO	IC-CO-1999-369680-hum	YES	NO	
IC	PHO1275			sm1,V1?	1962	hum	Guajira, Venezuela	VE	IC-VE-1962-PHO1275-hum	YES	NO	
IC	12 225			V2	1995	hum	Venezuela	VE	IC-VE-1995-12 225-hum	YES	NO	
IC	12 399			u,BHK1	1995	hum	Venezuela	VE	IC-VE-1995-12 399-hum	YES	NO	
IC	INH9813			u	1995	hum	Urdueta, Lara State, Venezuela	VE	IC-VE-1995-INH9813-hum	YES	NO	
IC	243938			V1,BHK1	1996	hor	Trujillo State, Venezuela	VE	IC-VE-1996-243938-hor	YES	NO	serum
IC	125567			BHK1	1997	hum	Zulia State, Venezuela	VE	IC-VE-1997-125567-hum	YES	NO	
IC	SH5			V1	1997	hum	Candelaria, Venezuela	VE	IC-VE-1997-SH5-hum	YES	NO	
IC	6803			V1	1999	hum	Falcon State, Venezuela	VE	IC-VE-1999-6803-hum	YES	NO	
IC	9813			V1	1999	hum	Urdueta, Lara State, Venezuela	VE	IC-VE-1999-9813-hum	YES	NO	
IC	ZGH734			V1	1999	hum	Sinamaica, Venezuela	VE	IC-VE-1999-ZGH734-hum	YES	NO	
IC	ZGH868			V1	1999	hum	Sinamaica, Venezuela	VE	IC-VE-1999-ZGH868-hum	YES	NO	
IC	255005			SM3	2000	hor	Barinas State, Venezuela	VE	IC-VE-2000-255005-hor	YES	NO	
IC	255058			sm2,V1	2000	hor	Carabobo State, Venezuela	VE	IC-VE-2000-255058-hor	YES	NO	
IC	12 563			V1,BHK1	u	hum	Venezuela	VE	IC-VE-u-12 563-hum	YES	NO	
IC	25716			BHK1	u	u	Venezuela	VE	IC-VE-u-25716-u	YES	NO	
IC	25717			BHK1	u	u	Venezuela	VE	IC-VE-u-25717-u	YES	NO	
ID	312714			V2	1978	rat	Pto. Boyaca, Colombia	CO	ID-CO-1978-312714-rat	YES	NO	Proechimys sp.
ID	92CO-59			none	1996	ham	Los Corales, Colombia	CO	ID-CO-1996-92CO-59-ham	YES	NO	brain
ID	980027			V1	1998	ham	Bosque San Miguel, Colombia	CO	ID-CO-1998-980027-ham	YES	NO	spleen
ID	980408			V1	1998	mos	Casanare, Colombia	CO	ID-CO-1998-980408-mos	YES	NO	Culex sp.
ID	005MH290			none	2000	ham	Monte San Miguel, Colombia	CO	ID-CO-2000-005MH290-ham	YES	NO	
ID	005MM515-11C			none	2000	mos	Monte San Miguel, Colombia	CO	ID-CO-2000-005MM515-11C-mos	YES	NO	
ID	622-41			none	2000	mos	Monte San Miguel, Colombia	CO	ID-CO-2000-622-41-mos	YES	NO	

ID	98-003	u	2002	ham	Los Corales, Colombia	CO	ID-CO-2002-98-003-ham	YES	NO	
ID	98-007	u	2002	ham	Los Corales, Colombia	CO	ID-CO-2002-98-007-ham	YES	NO	
ID	980019	V1	2002	ham	Bosque San Miguel, Colombia	CO	ID-CO-2002-980019-ham	YES	NO	heart
ID	980267	V1	2002	ham	Puerto boyaca, Colombia	CO	ID-CO-2002-980267-ham	YES	NO	heart
ID	980517	V1	2003	mos	San Pedro de la Paz, Colombia	CO	ID-CO-2003-980517-mos	YES	NO	Aedes angustivittatus
ID	247168	V2	2010	hor	Panama	PA	ID-PA-2010-247168-hor	YES	NO	brain
ID	247186	V2	2010	hor	Panama	PA	ID-PA-2010-247186-hor	YES	NO	brain
ID	FSL2314	u	2006	u	Loreto, Peru	PE	ID-PE-2006-FSL2314-u	YES	NO	
ID	FSL2649	u	2006	u	Loreto, Peru	PE	ID-PE-2006-FSL2649-u	YES	NO	
ID	FVB0204	u	2006	u	Cochabamba, Peru	PE	ID-PE-2006-FVB0204-u	YES	NO	
ID	FVB0258	u	2007	u	Cochabamba, Peru	PE	ID-PE-2007-FVB0258-u	YES	NO	
ID	FPI3700	V1	u	hum	Peru	PE	ID-PE-u-FPI3700-hum	YES	NO	
ID	249443	sm2,V6	1972	ham	Yumare, Venezuela	VE	ID-VE-1972-249443-ham	YES	NO	
ID	Pan34958	p2,sm1	1976	mos	Catatumbo, Zulia State, Venezuela	VE	ID-VE-1976-Pan34958-mos	YES	NO	Cx. ferreri?
ID	ZPC10	V1	1997	ham	Venezuela	VE	ID-VE-1997-ZPC10-ham	YES	NO	serum
ID	ZPC820	none	1997	ham	Las Nubes, catatumbo, Venezuela	VE	ID-VE-1997-ZPC820-ham	YES	NO	
ID	MAC10	V1	1998	ham	Padron Agric. Station, Miranda	VE	ID-VE-1998-MAC10-ham	YES	NO	
IE	68U201	u,sm1	1968	ham	La Avellana, Guatemala	GA	IE-GA-1968-68U201-ham	YES	NO	
III	PC254	u	1997	rat	Iquitos, Peru		III--1997-PC254-Proechimys spp	YES	NO	Proechimys spp
III	PE407660	u	1998	mos	Iquitos, Peru	PE	III-PE-1998-PE407660-mos	YES	NO	
III	FSL0190	V2	2000	hum	San Juan, Iquitos, Peru	PE	III-PE-2000-FSL0190-hum	YES	NO	

Array isolate	Closest genome	Closest to correct genome?	Allele differences from closest genome	Concordant calls	Total calls	Allele differences from correct genome	Percent concordant
IAB-CO-1967-CoAn5384-hor	IAB-CO-1967-CoAn5384-hor	Yes	28	1993	2021	28	98.6
IAB-GA-1969-69Z1-hum	IAB-GA-1969-69Z1-hum	Yes	32	1987	2019	32	98.4
IAB-PE-1973-111/73-hor	IAB-PE-1973-111/73-hor	Yes	39	1959	1998	39	98.0
IAB-VE-1938-Beck Wycoff-hor	IAB-VE-1938-Beck_Wycoff-hor	Yes	60	1971	2031	60	97.0
IC-CO-1961-V178-hor	IC-CO-1961-V178-hor	Yes	73	2132	2205	73	96.7
IC-CO-1962-V198-hum	IC-CO-1962-VEU55342-hum	Yes	38	2093	2131	38	98.2
IC-VE-1962-PHO127-hum	IC-VE-1962-PHO127-hum	Yes	59	2072	2131	59	97.2
IC-VE-1963-P676Ag-mos	ID-VE-1963-AF375051-mos	Yes	37	2089	2126	37	98.3
IC-VE-1964-PMCHo5-hum	IC-VE-1964-VEU55345 PMCHo5-hum	Yes	47	2080	2127	47	97.8
IC-VE-1992-243937-hor	IC-VE-1992-243937-hor	Yes	26	2060	2086	26	98.8
IC-VE-1993-SH3-hum	IC-VE-1993-SH3-hum	Yes	42	2035	2077	42	98.0
IC-VE-1995-6119-hum	IC-VE-1995-VEU55347 6119-hum	Yes	44	2086	2130	44	97.9
IC-VE-2000-254934-hor	IC-VE-2000-255010-hor	No	43	2080	2124	44	97.9
IC-VE-2000-255010-hor	IC-VE-2000-255010-hor	Yes	37	2088	2125	37	98.3
ID-CO-1960-V209A-mus	ID-CO-1960-V209A-mus	Yes	58	2139	2197	58	97.4
ID-CO-1969-CoAn9004-ham	IF-BR-1978-78V-3531-u	No	291	1010	1693	683	59.7
ID-CO-1971-307537-mos	ID-CO-1971-307537-mos	Yes	50	2160	2210	50	97.7
ID-CO-1973-309506-ham	ID-CO-1973-309506-ham	Yes	69	2132	2201	69	96.9
ID-CO-1974-309752-hum	ID-CO-1974-309572-hum	Yes	55	2138	2193	55	97.5
ID-CO-1977-334250-mos	ID-CO-1977-334250-mos	Yes	50	2139	2189	50	97.7
ID-CO-1978-335733-ham	ID-CO-1978-335733-ham	Yes	67	2115	2182	67	96.9
ID-CO-1997-97Co42-ham	**ID-CO-1997-97Co42-ham**	**Yes**	**41**	**2094**	**2135**	**41**	**98.1**
ID-CO-1997-97Co42-ham	**ID-CO-1997-97Co42-ham**	**Yes**	**68**	**2067**	**2135**	**68**	**96.8**
ID-CO-1999-99SMM304-1-mos	ID-CO-1999-993MM304-1-mos	Yes	50	2067	2117	50	97.6
ID-CO-2000-00SMH264-ham	ID-CO-2000-00SMH264-ham	Yes	76	2074	2150	76	96.5
ID-CO-u-CoAn9145-ham	ID-CO-u-CoAn59145-ham	Yes	45	2164	2209	45	98.0
ID-EC-1975-76V2561-mos	IF-BR-1978-78V-3531-u	No	293	972	1655	683	58.7
ID-EC-1976-R16880-ham	ID-VE-1972-249443-ham	No	289	986	1712	726	57.6
ID-PA-1961-3880-hum	ID-PA-1961-3880-hum	Yes	53	1967	2020	53	97.4
ID-PA-1961-4840-hum	ID-PA-1961-4840-hum	Yes	57	2019	2076	57	97.3
ID-PA-1962-8138-hum	**ID-PA-1962-8138-hum**	**Yes**	**49**	**2020**	**2069**	**49**	**97.6**
ID-PA-1962-8138-hum	**ID-PA-1962-8138-hum**	**Yes**	**22**	**2047**	**2069**	**22**	**98.9**
ID-PA-1966-242959-u	ID-PA-1966-242959-u	Yes	51	2019	2070	51	97.5
ID-PA-1977-903104-mos	ID-PA-1977-903104-mos	Yes	53	1959	2012	53	97.4
ID-PA-1984-GML903843-hum	ID-PA-1984-GML903843-hum	Yes	62	1914	1976	62	96.9
ID-PA-1997-474590-hum	**ID-PA-1997-474590-hum**	**Yes**	**34**	**1882**	**1916**	**34**	**98.2**
ID-PA-1997-474590-hum	**ID-PA-1997-474590-hum**	**Yes**	**38**	**1878**	**1916**	**38**	**98.0**
ID-PA-2000-481460-u	ID-PA-2000-481460-u	Yes	54	1902	1956	54	97.2
ID-PA-2001-484551-u	ID-PA-2001-484551-u	Yes	60	1925	1985	60	97.0
ID-PA-2001-485029-u	ID-PA-2001-485029-u	Yes	55	1922	1977	55	97.2
ID-PA-2003-212857-hum	ID-PA-2003-212857-hum	Yes	46	1934	1980	46	97.7
ID-PA-2003-213391-hum	ID-PA-2003-213391-hum	Yes	27	1760	1787	27	98.5
ID-PE-1975-75D143-mos	ID-PE-1975-75D143-mos	Yes	48	2130	2178	48	97.8
ID-PE-1995-IQT1724-hum	ID-PE-1995-IQT1724-hum	Yes	44	1929	1973	44	97.8
ID-PE-1998-02-2720-98-hum	ID-PE-1998-02_2720_98-hum	Yes	43	1913	1956	43	97.8
ID-PE-2000-FSE507-hum	ID-PE-2000-FSE507-hum	Yes	51	2060	2111	51	97.6
ID-VE-1973-204381-mos	ID-VE-1973-204381-mos	Yes	64	2063	2127	64	97.0
ID-VE-1974-23647-ham	ID-VE-1974-23647-ham	Yes	48	2150	2198	48	97.8
ID-VE-1976-251641-ham	ID-VE-1976-251641-ham	Yes	70	2146	2216	70	96.8
ID-VE-1997-ZPC727 v2-ham	ID-VE-1997-ZPC727-ham	Yes	47	2027	2074	47	97.7
IE-BE-1967-67U201-ham	IE-BE-1967-67U201-ham	Yes	46	1751	1797	46	97.4
IE-GA-1968-68U200-ham	IE-GA-1968-68U200-ham	Yes	57	1709	1766	57	96.8
IE-GA-1968-68U217-ham	IE-GA-1968-68U217-ham	Yes	48	1723	1771	48	97.3
IE-GA-1970-70U74-ham	IE-GA-1970-70U74-ham	Yes	47	1729	1776	47	97.4
IE-GA-1971-71U382-ham	IE-GA-1971-71U382-ham	Yes	53	1730	1783	53	97.0
IE-GA-1971-71U384-ham	IE-GA-1971-71U384-ham	Yes	68	1710	1778	68	96.2
IE-GA-1972-72U23-ham	IE-GA-1972-72U23-ham	Yes	32	1734	1766	32	98.2
IE-GA-1973-73U151-ham	IE-GA-1973-73U151-ham	Yes	43	1718	1761	43	97.6
IE-GA-1978-78U202-ham	IE-GA-1978-78U202-ham	Yes	43	1728	1771	43	97.6
IE-GA-1979-79U13-ham	IE-GA-1979-79U13-ham	Yes	47	1717	1764	47	97.3
IE-GA-1980-80U76-hor	IE-GA-1980-80U76-hor	Yes	27	1733	1760	27	98.5
IE-HO-1967-67U208-ham	IE-HO-1967-67U208-ham	Yes	42	1703	1745	42	97.6
IE-HO-1967-67U225-ham	IE-HO-1967-67U225-ham	Yes	61	1697	1758	61	96.5
IE-ME-1963-63A216-mos	IE-ME-1963-63A216-mos	Yes	51	1747	1798	51	97.2
IE-ME-1963-63Z1-hum	IE-ME-1963-63Z1-hum	Yes	47	1753	1800	47	97.4
IE-ME-1965-65U206-ham	IE-ME-1965-65U206-ham	Yes	32	1772	1804	32	98.2
IE-ME-1966-66U91-ham	IE-ME-1966-66U91-ham	Yes	40	1780	1820	40	97.8
IE-ME-1967-67U222-ham	IE-ME-1967-67U222-ham	Yes	53	1698	1751	53	97.0
IE-ME-1969-68U315-ham	IE-ME-1969-69U315-ham	Yes	47	1741	1788	47	97.4
IE-ME-2001-MX01-32-ham	IE-ME-2001-MX01-32-ham	Yes	96	1647	1743	96	94.5
IE-ME-2003-MX03-H1-ham	IE-ME-2003-MX03H2-ham	No	41	1670	1713	43	97.5
IE-ME-2003-MX03-H2-ham	IE-ME-2003-MX03H2-ham	Yes	14	1700	1714	14	99.2
IE-ME-2008-MX08-H50-ham	IE-ME-2008-MX08H50-ham	Yes	50	1699	1749	50	97.1
IE-ME-2008-MX08-H53-ham	IE-ME-2008-MX08H53-ham	Yes	54	1723	1777	54	97.0
IE-ME-2008-MX09-M64-mos	IE-ME-2009-MX09M51-mos	No	51	1702	1754	52	97.0
IE-ME-2009-MX09-Eq03-hor	IE-ME-2009-MX09Eq03-hor	Yes	23	1722	1745	23	98.7
IE-ME-2010-MX10-H91-ham	IE-ME-2010-MX10_91M8-mos	No	47	1690	1738	48	97.2
IE-NI-1968-2177B-u	IE-NI-1968-2177B-u	Yes	40	1752	1792	40	97.8
IE-PA-1961-BT2607-mos	IE-PA-1961-BT2607-mos	Yes	23	1442	1465	23	98.4
IE-PE-1970-70U1134-ham	IE-PE-1970-70U1134-ham	Yes	45	2147	2192	45	97.9
IE-VE-1962-Menall-hum	IE-VE-1962-Mena II-hum	Yes	14	1443	1457	14	99.0
IIIB-BR-1954-BeAn8-u	IIIB-BR-1954-Mucambo BeAn 8-u	Yes	3	768	771	3	99.6
IV-BR-1961-BeAr 35645-mos	IF-BR-1978-78V-3531-u	No	295	375	693	318	54.1
V-FG-1968-CABV AR508-mos	V-FG-1968-Cabassou CaAr 508-mos	Yes	1	617	618	1	99.8
Totals				153518	159629		96.2

Virus strain	Experimental treatment	Replicate mouse number	Closest genome	Allele differences from closest genome	Concordant calls	Total calls	Allele differences from correct genome	Percent concordant
VEEV IAB-VE-vaccine-TC83	Extract from brain of infected mouse	1	IAB-VE-vaccine-TC-83-u	45	1969	2014	45	97.8
		2	IAB-VE-vaccine-TC-83-u	38	1976	2014	38	98.1
		3	IAB-VE-vaccine-TC-83-u	49	1965	2014	49	97.6

(*) Probe signals on this array were close to background level for all probes.